# STRATEGIES FOR ABSOLUTE QUANTIFICATION IN PROTEOMICS

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by Jennifer Rivers

February 2008

## ACKNOWLEDGEMENTS

I would like to thank all members of the Proteomics and Functional Genomics Group, Veterinary Preclinical Sciences for their help, advice and support for the duration of my post-graduate studies. I have enjoyed my time here immensely and have become very much part of 'the family'. In particular, I thank my supervisor Rob Beynon for continued support and encouragement in addition to exceptional training and guidance to facilitate my development as a research scientist.

Special thanks go to all of those involved in the original development of the QconCAT strategy, particularly Simon Gaskell and Julie Pratt who have supported the progress of the QconCAT method facilitating the publication of relevant work. For expression and labelling of QconCAT proteins, in addition to constant support and a friendly ear, I would specifically like to thank Deb Simpson. I would also like to thank Duncan Robertson for training in instrumentation, particularly mass spectrometry, and Lynn McLean for lab-based training, providing a solid foundation for my experimental work.

I would like to thank the Roslin Institute, Edinburgh, UK for growing birds and collecting muscle samples, particularly Heather McCormack. I am also grateful to Ian Edwards for the benefit of his experience working with chicken skeletal muscle, in addition to sharing quantitative data for comparison. I have also compared quantitative data with that obtained by Julia Hayter whom also merits special thanks.

For funding my PhD, I am grateful to the BBSRC and Genus plc. My supervisors at Genus, Dominque Rocha and Gary Evans have shown a continued interest in my work and this has provided me with considerable opportunity to participate in formal presentation throughout my postgraduate education, in addition to well designed training courses covering several disciplines essential to a career in science, and for this I am extremely grateful.

I have enjoyed collaborating with Daniel Coca and Istvan Bogdan at the University of Sheffield to investigate superior strategies for mass spectral data processing, resulting in two published outputs. This has deviated from my main line of study and has been both enjoyable and informative.

For investigation of the post-translational modification deamidation, I am grateful to the efforts of Ian Edwards who initially observed the modification, and Lucy McDonald who conducted initial investigation of this process. This work developed over a considerable period of time, and its final publication in January 2008 was a significant achievement (and somewhat of a relief!).

Lastly, but by no means least, special thanks go to my family and friends for continued love and support, particularly over the last few months. I would not have achieved to this level without their unwavering belief in my abilities and a desire to see me succeed.

"Qualitative data is important, Quantitative data is becoming essential"

Professor S.J. Gaskell, 2006

## TABLE OF CONTENTS

**List of figures and tables** ..... i

**Synopsis** ..... v

**1. Introduction to Quantitative Proteomics** ..... 1

1.1 'Omics' technologies and the proteome ..... 1
1.1.1 Why proteomics? ..... 1
1.1.2 The nature of proteomics ..... 2

1.2 Stable isotopes and their significance for proteomics ..... 6
1.2.1 Isotope abundance ..... 7
1.2.2 The ability of MS to resolve isotopically labelled peptides ..... 7
1.2.3 Mass isotopomer distribution analysis ..... 8
1.2.4 Quantitative proteomics ..... 9

1.3 Protein Identification ..... 12
1.3.1 Gel electrophoresis ..... 12
1.3.2 Mass spectrometry ..... 12
1.3.3 Peptide mass fingerprinting ..... 16
1.3.4 Tandem mass spectrometry ..... 17
1.3.5 MSMS ion searching for protein identification ..... 19

1.4 Challenges in peptide based proteomics ..... 21
1.4.1 Proteolysis ..... 21
1.4.2 Sample complexity and dynamic range ..... 24

1.5 Quantification Proteomics ..... 29
1.5.1 The significance of quantification for proteomics ..... 29
1.5.2 Relative quantification ..... 30
1.5.3 Absolute quantification ..... 39
1.5.4 Challenges for absolute quantification using surrogate peptides ..... 42

**2. Aims and Objectives** ..... 44

**3. The QconCAT strategy for absolute quantification of chicken skeletal muscle soluble proteins** ..... 44

3.1 Changing proteome dynamics in chicken skeletal muscle ..... 44

3.2 QconCAT design ..... 45

3.3 Deployment of the QconCAT strategy ..... 46
3.3.1 Reliability of a QconCAT method ..... 46
3.3.2 Proteolysis ..... 47

3.4 Additional applications of QconCAT ..... 47
3.4.1 Using QconCAT to quantify skeletal muscle proteins from other species ..... 47
3.4.2 Quantification of normalisation using Equalizer™ beads ..... 47
3.4.3 Absolute quantification of the post-translation modification, deamidation ..... 48

**4. Materials and Methods** ..... 50

**5. Results and Discussion ..... 58**

5.1 Design, preparation, purification and analysis of QconCAT and analyte proteins ..... 58
5.1.1 Design, preparation and purification of QconCAT ..... 58
5.1.2 Proteomic analysis of QconCAT ..... 59
5.1.3 Proteomic analysis of chicken skeletal muscle soluble proteins ..... 63

5.2 Proteolysis of QconCAT and analyte proteins ..... 63
5.2.1 Proteolysis of QconCAT ..... 63
5.2.2 Proteolysis of chicken skeletal muscle soluble proteins ..... 65

5.3 Sample complexity and dynamic range ..... 67
5.3.1 Mass spectrometry for absolute quantification using the QconCAT method ..... 67
5.3.2 Challenges for data acquisition and analysis for quantification ..... 68

5.4 Validation of the QconCAT method ..... 71
5.4.1 Quantification of unlabelled QconCAT by labelled QconCAT ..... 71
5.4.2 Variance in the QconCAT method ..... 71
5.4.3 Accuracy of the QconCAT method ..... 72
5.4.4 Comparison of the QconCAT method with alternative strategies for absolute quantification ..... 72

5.5 Absolute quantification of chicken skeletal muscle soluble proteins ..... 75

5.6 Additional applications of QconCAT technology ..... 78
5.6.1 Quantification of soluble skeletal muscle proteins in other species ..... 78
5.6.2 Quantification of normalisation using Equalizer™ bead technology ..... 79
5.6.3 Absolute quantification of the post-translational modification, deamidation ..... 82

**6. Conclusions ..... 89**

**7. References ..... 98**

**8. Supplementary figures ..... 110**

**9. Publications as a consequence of this thesis ..... 135**

## FIGURES

Following page

Figure 1. Stable isotope labelling of a tryptic peptide ..... 7
Figure 2. Ionisation by MALDI and ESI for mass spectrometry ..... 13
Figure 3. Resolution of peptide mass spectra ..... 14
Figure 4. Time of flight and quadrupole mass analysers ..... 14
Figure 5. Quadrupole ion trap and ion cyclotron resonance ..... 15
Figure 6. Fragmentation of a peptide ion ..... 18
Figure 7. Nomenclature of proteolysis according to Schechter and Berger (1967) ..... 24
Figure 8. Stable isotope labelling strategies for comparative proteomics ..... 34
Figure 9. Isotope coded affinity tagging (ICAT) ..... 35
Figure 10. Isobaric tagging for relative and absolute quantitation (iTRAQ) ..... 36
Figure 11. Stable isotope labelling by amino acids in cell culture (SILAC) ..... 37
Figure 12. Internal standardisation for absolute protein quantification ..... 40
Figure 13. The QconCAT strategy for absolute quantification ..... 41
Figure 14. Growth rates of broiler and layer chickens ..... 44
Figure 15. Selection of QconCAT peptides and subsequent design of QconCAT gene constructed for the absolute quantification of chicken skeletal muscle soluble proteins ..... 45
Figure 16. QconCAT expression and purification ..... 58
Figure 17. ESI-Q-ToF MS analysis of unlabelled QconCAT protein ..... 59
Figure 18. QconCAT protein digested in-solution with trypsin ..... 59
Figure 19. Confirmation of pyro-glutamic acid modification to QconCAT peptides ..... 59
Figure 20. Distinction between unlabelled and [$^{15}$N] labelled QconCAT peptides in MALDI-ToF mass spectra . 59
Figure 21. Diagnostic peptide mass fingerprinting to distinguish between two major protein bands upon SDS-PAGE separation of QconCAT protein ..... 60
Figure 22. Carbamylation of QconCAT peptides ..... 60
Figure 23. Implementation of strategies to de-salt QconCAT protein preparations ..... 60
Figure 24. Purification of QconCAT labelled (H) and unlabelled (L) using HisTrap™ columns without prior solubilisation in urea ..... 61
Figure 25. Purification of unlabelled QconCAT protein, solubilised in 6M guanidinium chloride ..... 61
Figure 26. Purification of [$^{13}C_6$]lys, [$^{13}C_6$]arg-labelled QconCAT protein, solubilised using 8M urea or 6M guanidinium chloride ..... 62
Figure 27. Purification of [$^{13}C_6$]lys, [$^{13}C_6$]arg-labelled QconCAT, solubilised in 6M guanidinium chloride and de-salted by dialysis ..... 62
Figure 28. Unlabelled QconCAT protein purified in 6M guanidinium chloride ..... 62
Figure 29. Disinction between unlabelled and [$^{13}C_6$]lys, [$^{13}C_6$]arg-labelled QconCAT peptides in MALDI-ToF mass spectra ..... 62
Figure 30. Isotope distribution of a Q-peptide labelled with [$^{13}C_6$]lys/arg and [$^{15}$N] in MALDI-ToF mass spectra ..... 62

Figure 31. Chicken skeletal muscle soluble protein expression in broiler and layer birds over 30 days of growth ..... 63
Figure 32. Chicken skeletal muscle soluble proteins digested in-solution with trypsin ..... 63
Figure 33. Proteolysis of QconCAT with trypsin ..... 64
Figure 34. Proteolysis of QconCAT and diagnostic peptide mass fingerprinting of peptides ..... 64
Figure 35. Proteolysis of QconCAT and extensive separation of protein fragments by 1D SDS-PAGE ..... 64
Figure 36. Proteolysis of QconCAT and analysis of peptide release ..... 64
Figure 37. In-solution proteolysis of chicken skeletal muscle soluble proteins with trypsin and the effect of acetonitrile ..... 65
Figure 38. In-solution proteolysis of chicken skeletal muscle soluble proteins with trypsin, the effect of acetonitrile, and protein denaturation prior to addition of protease ..... 65
Figure 39. Absolute quantification of unlabelled QconCAT proteolysis ..... 66
Figure 40. Quantification of 500min proteolysis of analyte proteins with trypsin using QconCAT ..... 66
Figure 41. Quantification of 30h proteolysis of analyte proteins with trypsin using QconCAT ..... 66
Figure 42. Absolute quantification using MALDI-ToF MS or LC-ESI-Q-ToF MS ..... 67
Figure 43. Relative intensity of QconCAT labelled and unlabelled peptides acquired from 15 different locations on a MALDI target ..... 67
Figure 44. Heavy:light ion pairs for analyte protein quantification in MALDI-ToF MS ..... 68
Figure 45. Guanidination of QconCAT peptides ..... 69
Figure 46. Relative signal intensity of guanidinated QconCAT peptides ..... 69
Figure 47. Guanidination of chicken skeletal muscle soluble peptides ..... 69
Figure 48. Chromatographic elution of analyte and internal standard peptides ..... 69
Figure 49. Isolation of analyte:standard peptide pairs by reversed-phase chromatography ..... 70
Figure 50. Fractionation of QconCAT peptides by LC-MALDI-ToF MS ..... 70
Figure 51. Quantification of chicken skeletal muscle soluble proteins by LC-MALDI-ToF MS ..... 70
Figure 52. Consistency of quantification using LC-MALDI-ToF MS and MALDI-ToF MS ..... 71
Figure 53. Validation of quantification using a mixture of unlabelled and labelled QconCAT proteins ..... 71
Figure 54. Sources of variance in a QconCAT experiment ..... 71
Figure 55. Sources of variance in a QconCAT experiment ..... 71
Figure 56. Relationship between coefficient of variance and ion signal intensity in MALDI-ToF mass spectra ..... 72
Figure 57. Accuracy of the QconCAT method using purified proteins ..... 72
Figure 58. Accuracy of quantification using QconCAT ..... 72
Figure 59. Quantification of a synthetic peptide internal standard using QconCAT ..... 73
Figure 60. Quantification of a synthetic peptide internal standard stored under different conditions, using QconCAT ..... 73
Figure 61. Comparison of QconCAT and synthetic peptide for quantification ..... 73
Figure 62. Comparison of protein quantification by densitometry and QconCAT (broilers) ..... 74
Figure 63. Comparison of protein quantification by densitometry and QconCAT (layers) ..... 74
Figure 64. Comparison of protein quantification by densitometry and QconCAT ..... 74
Figure 65. Quantification of increasing amount of total protein, by densitometry and the QconCAT method ..... 74

Figure 66. Quantification of increasing amount of total protein, by densitometry and the QconCAT method ..... 74
Figure 67. Comparison of quantification using intact mass analysis by ESI-Q-ToF MS and the QconCAT method ..... 75
Figure 68. Comparison of quantification using intact mass analysis by ESI-Q-ToF MS, SDS-PAGE and densitometry analysis, and the QconCAT method ..... 75
Figure 69. SDS-PAGE analysis of QconCAT and chicken skeletal muscle soluble proteins from broiler and layer strains ..... 76
Figure 70. Quantification of GAPDH expression in chicken skeletal muscle ..... 76
Figure 71. Quantification of chicken skeletal muscle protein expression by QconCAT ..... 76
Figure 72. Comparison of absolute protein quantification of soluble skeletal muscle proteins in broiler and layer chickens ..... 77
Figure 73. Quantification of skeletal muscle proteins from rabbit and chicken by QconCAT ..... 78
Figure 74. Skeletal muscle soluble proteins from carp, mouse and chicken ..... 78
Figure 75. Analyte: internal standard peptide pairs for cross-species quantification ..... 78
Figure 76. Normalisation of chicken skeletal muscle soluble protein abundance using Equalizer™ beads ..... 80
Figure 77. Densitometry analysis of 1D SDS-PAGE separated chicken skeletal muscle soluble proteins normalised using Equalizer™ beads ..... 80
Figure 78. Identification of normalised proteins using in-gel digestion with trypsin and LC-ESI-LTQ MSMS ..... 80
Figure 79. Proteolysis of chicken skeletal muscle soluble proteins and normalised material ..... 81
Figure 80. Quantification of normalisation of chicken skeletal muscle soluble proteins using QconCAT ..... 81
Figure 81. Analyte and internal standard peptide pairs for two proteins normalised using Equalizer™ beads ... 81
Figure 82. Chromatographic confirmation of increase in abundance of API with normalisation using Equalizer™ beads ..... 81
Figure 83. Confirmation of increase in abundance of API with normalisation using Equalizer™ beads ..... 81
Figure 84. Atypical peptide mass spectrum consistent with deamidation ..... 82
Figure 85. Esterification of acidic residues in the N-terminal peptide of GAPDH ..... 83
Figure 86. Time course of deamidation of the N-terminal peptide of GAPDH ..... 83
Figure 87. Time course of deamidation of the N-terminal GAPDH during proteolysis with Asp-N ..... 83
Figure 88. Ion signal response in MALDI-ToF MS from asparagine and aspartic acid containing peptide ..... 84
Figure 89. The effect of temperature on peptide deamidation ..... 84
Figure 90. Time course of deamidation of the N-terminal peptide of GAPDH ..... 85
Figure 91. Model of proteolysis and deamidation of the N-terminal peptide of GAPDH ..... 85
Figure 92. Absolute quantification of proteolysis of the GAPDH N-terminus ..... 86
Figure 93. 3D structure of rabbit skeletal muscle GAPDH ..... 86
Figure 94. Proteolysis of GAPDH with trypsin ..... 86
Figure 95. The effect of denaturing protein structure by heating on the rate of deamidation ..... 87
Figure 96. Absolute quantification of GAPDH based on a deamidating peptide containing an internal trypsin cleavage site ..... 88

## TABLES

Table 1. Natural abundance of stable isotopes most commonly used in proteomics ..... 7
Table 2. Commonly used proteases for proteomics ..... 21
Table 3. Peptides selected from chicken skeletal muscle soluble proteins represented in the QconCAT protein ..... 45
Table 4. Identification of chicken skeletal muscle proteins by peptide mass fingerprinting ..... 63
Table 5. Identification of normalised proteins using in-gel digestion with trypsin and LC-ESI- LTQ-MSMS ..... 80

## STRATEGIES FOR ABSOLUTE QUANTIFICATION IN PROTEOMICS

The protein composition of a biological entity; its proteome, is the subject of study in proteomics. Proteomics in turn can be conceptually subdivided into the identification of proteins present, characterisation of function and interactions with other proteins, and protein quantification. As an example, proteins associated with the onset and progression of disease, once identified and characterised may be quantified, detecting minor fluctuations in abundance to facilitate diagnosis and treatment. For systems biology and global comparisons across multiple platforms, absolute protein quantification is preferred, rather than measuring the abundance of proteins relative to a second cellular state. To achieve this, techniques based on well established precepts of stable isotope dilution have been developed. These involve the incorporation of naturally occurring stable isotopes into differential labels to compare protein abundance of two or more samples, or internal standards added in known amounts. Proteolytic digestion of differentially labelled samples creates a peptide that elicits a known mass shift on mass spectrometric analysis. Using the peptide as a surrogate for the protein of interest, the signal intensity of unlabelled and labelled ions can be reconciled to measure relative abundance. For absolute protein quantification, stable isotopes are incorporated into synthetic proteotypic peptides for use as internal standards designed to mimic native peptides formed by proteolysis. However, for the absolute quantification of several proteins in a biological system, a stable isotope labelled peptide would have to be synthesised, at relatively high cost, for each protein to be quantified.

To create a multiplexed method for protein quantification, *de novo* gene design has been used to create and express artificial proteins (QconCATs) that comprise a concatenation of proteotypic peptides. Upon complete proteolytic digestion, each peptide was produced in equimolar amounts, permitting absolute quantification of multiple proteins in a single experiment. One QconCAT protein contained a tryptic peptide from each of twenty proteins present in the soluble fraction of chicken skeletal muscle. Optimised DNA sequences encoding these peptides were concatenated and inserted into a vector for high level expression in *E.coli*. The protein was expressed in a minimal medium enriched with stable isotopes, or containing selectively labelled amino acids, creating an equimolar series of uniformly labelled proteotypic peptides. The labelled QconCAT protein, purified by affinity chromatography and quantified was added to a homogenised muscle preparation in a known amount prior to proteolytic digestion with trypsin.

The goal of this study was to define the deployment, sources of error and statistical behavior of a QconCAT analysis. Analytical challenges to exploring the proteome, particularly for absolute quantification were addressed including proteolysis of analyte and internal standard proteins, sample complexity, dynamic range, and ionisation in mass spectrometry. As anticipated, the QconCAT was completely digested at a rate far higher than the analyte proteins, confirming the applicability of such artificial proteins for multiplexed quantification. In addition, this demonstrates further application of readily digested QconCAT proteins to assessment of proteolysis kinetics in the analyte system. Alternative mass spectrometric approaches with and without prior reversed phase separation have been investigated, particularly LC-ESI-ToF MS and MALDI-ToF MS for analysis of tryptic peptides. This has established the use of QconCAT technology for absolute quantification using various methods of detection and analysis. Accuracy and reproducibility of the QconCAT method in addition to the nature of technical variance compared with biological variance have been assessed in a complete study involving six time points during growth with four birds at each time point for both broiler and layer strains. Absolute quantification using the QconCAT method was equivalent to alternative strategies tested, including the single synthetic peptide approach. This thesis concludes that QconCATs offer a new and efficient approach to precise and simultaneous absolute quantification of multiple proteins.

As an extension, additional applications of QconCAT technology have been investigated including the robustness for use to quantify the same proteins in other species, absolute quantification of post-translational modifications, and quantification of protein abundance normalisation during enrichment with peptide library beads. This highlights the diversity of QconCAT technology and its advantages over alternative strategies for absolute quantification.

## 1. Introduction to Quantitative Proteomics


1.1 'Omics' technologies and the proteome ..... 1
1.1.1 Why proteomics? ..... 1
1.1.2 The nature of proteomics ..... 2

1.2 Stable isotopes and their significance for proteomics ..... 6
1.2.1 Isotope abundance ..... 7
1.2.2 The ability of MS to resolve isotopically labelled peptides ..... 7
1.2.3 Mass isotopomer distribution analysis ..... 8
1.2.4 Quantitative proteomics ..... 9

1.3 Protein Identification ..... 12
1.3.1 Gel electrophoresis ..... 12
1.3.2 Mass spectrometry ..... 12
1.3.3 Peptide mass fingerprinting ..... 16
1.3.4 Tandem mass spectrometry ..... 17
1.3.5 MSMS ion searching for protein identification ..... 19

1.4 Challenges in peptide based proteomics ..... 21
1.4.1 Proteolysis ..... 21
1.4.2 Sample complexity and dynamic range ..... 24

1.5 Quantification Proteomics ..... 29
1.5.1 The significance of quantification for proteomics ..... 29
1.5.2 Relative quantification ..... 30
1.5.3 Absolute quantification ..... 39
1.5.4 Challenges for absolute quantification using surrogate peptides ..... 42

# 1. Introduction to Quantitative Proteomics

## 1.1 'Omics' Technologies and the Proteome

### 1.1.1 Why proteomics?

Dramatic advances in genomics, transcriptomics and proteomics in the 21st century have driven understanding of biological systems into the relatively new era of 'system biology'. Discoveries from all three disciplines contribute information in different forms, with different outcomes, but can be used in concert, providing a systems-wide viewpoint used to decipher the complexities and interactions of biological systems. First and foremost, genomics; the study of the entire genome of an organism, has revolutionised biological studies of different species, by sequencing their DNA, defining the number of genes, and mapping genetic variation, including the interaction between genes, and their influence on phenotype. Genome sequencing of entire organisms now takes place in the laboratory on a daily basis (for example Seedorf *et al.,* 2008), an achievement that could only have been dreamt of, even a few years ago. These analyses have prompted functional genomic studies, in which DNA/RNA microarrays are used to monitor changes in expression, or single nucleotide polymorphisms (SNPs) for single genes, or to profile thousands of genes simultaneously. SNPs can be identified with little or no prior knowledge of expressed phenotype, and can predict whether this polymorphism will have an effect on gene expression. To delve further into the biological 'meaning', profiling studies using mRNA expression; transcriptomics, have revealed that the products of expressed genes; proteins, are not necessarily expressed in strict correlation with the cognate mRNA, and also have uncharacterised functions, that cannot be predicted from the genome (Galperin and Koonin, 2004). This holds the potential for another dimension of information useful in a systems biology context and merits the study of the entire proteome of an organism, proteomics, which aims to characterise the function manifested by changes in gene expression in different cellular states (deHoog and Mann, 2004). Protein abundance depends not only on the levels of mRNA present, but also on events that occur during translation, for example alternative splicing or point mutations, and subsequent changes to the mature protein product including regulated degradation and protein stability through post-translational modifications or protein-protein interactions. Proteomics can measure the downstream consequences of cellular events, making the transition into 'systems biology', and combining data from changes in signalling events through transcription, translation and post-translational modification to metabolic

alterations (Julka and Regnier, 2004). Both transcriptomics and proteomics can provide quantitative expression data, although correlation between protein and mRNA abundance is weak (for example, Gygi *et al.,* 1999). However, simultaneous analysis can be used to predict the sources of discrepancy as these are two-fold; biological factors, such as alterations during, or after translation, or technical limitations (Nie *et al.,* 2006). The relationship between mRNA and protein abundance can be highly informative for example, the investigation of protein turnover kinetics, where protein synthesis is largely directed by mRNA expression, and protein degradation by the concentration of expressed protein in the cell or tissue (Yu *et al.,* 2007). This has prompted the development of a web based tool (http://proteomics.gersteinlab.org) in which datasets of both mRNA expression and protein abundance are available for comparison. However, these comparisons are currently limited, primarily by access to appropriate resources and often, expression analysis is undertaken at one level or the other. The advantages of studying mRNA expression include the requirement for less starting material due to amplification strategies, and faster analysis times for high throughput; expression profiling of thousands of genes simultaneously permit the transition of 'discovery' to 'browsing' mode where transcripts are identified and quantified in a single experiment (Aebersold, 2003). This is a long-term goal for proteomics in order to translate relevant discoveries to clinical diagnostics, but the leap in magnitude from genes (around 20,000-25,000 in the human genome; International human genome sequencing consortium, 2004) to proteins (in excess of 500,000 in human serum, Anderson and Anderson, 2002) provides a considerable challenge for current technology. Limitations of transcriptomics include questionable reproducibility across different platforms and between different laboratories, but most essentially that quantitative analysis of mRNA is not often reflected at the protein level. This is particularly observed for extracellular proteins, for example in blood and other body fluids, as mRNA measurements of a particular cell type or tissue are not relevant throughout the whole body. For clinical research, global quantitative proteomic profiling is essential for diagnosis and assessment of treatment regimens. As such, strategies for absolute quantification where protein abundance is expressed as number of molecules; the principal focus of this thesis, are in development, allowing systems-wide measurement of protein expression levels that is comparable to mRNA approaches (Cox and Mann, 2007).

### 1.1.2 The nature of proteomics

There are fundamentally three 'sub-types' of proteomic analysis; identification, characterisation and quantification. It is vital to know what the proteins are, discover what they do, how they

interact with each other and their environment, and finally to measure the abundance of particular proteins and how this may change and influence other processes, for example interactions with other proteins.

***Identification***

The molecular weight of native proteins can be determined by gel-based separation methods, or direct mass spectrometry (MS) analysis. However, for protein identification, analysis of molecular weight is not sufficient and strategies must delve further into the protein structure to ascertain the sequence of individual amino acids. Consequently, analysis of individual proteins and complex mixtures of proteins in biological systems is driven by two different philosophies; **'top down'** proteomics in which intact proteins are analysed by MS and fragmented directly, as individual proteins or in complex mixtures, and **'bottom up'** proteomics in which proteins are first digested into peptides, using a protease, for example trypsin, prior to analysis by MS and fragmentation to determine amino acid sequences. For top-down proteomics, intact proteins must be analysed directly; this has previously involved molecular weight determination using gel based methods which have limited dynamic and molecular weight range, but could permit ambiguous assignments of post-translational modifications in addition to a measure of protein purity. For the more sensitive and informative technique of MS, development of soft ionisation techniques in the 1980's, in which samples are transferred into the gas phase without extensive fragmentation, allowed proteins and peptides to be analysed in this way. This form of ionisation permitted transfer of large molecules into the gas phase without affecting their integrity. For intact protein analysis with fragmentation for protein identification and characterisation, proteins separated by gel electrophoresis, or alternative strategies requiring solubilisation of protein mixtures in-solution, are not compatible. By contrast, peptides are soluble and simple to separate prior to direct MS analysis, but provide an extra dimension of complexity. Additionally, digesting a complex mixture of proteins into peptides loses connectivity to the parent proteins, particularly important to ascertain the location of post-translational modifications on specific proteins. However, despite this, protein identification has become predominantly peptide based over the last 20 years (Cañas *et al.,* 2006). Instrumentation and strategies for MS analysis of intact proteins are also developing, and are becoming widely used in the proteomics community. The ability to fragment intact proteins provides excellent opportunities for analysis of post-translational modifications and will be discussed in the context of protein characterisation. For peptide-based protein identification, following MS analysis of proteolysed proteins, peptide masses are used to create a 'fingerprint' of the protein that they have been

digested from. These data are used to search databases of theoretical enzymatic digests of proteins that have previously been identified and for which the amino acid sequences are known. The success of this technique will depend on the mass accuracy delivered by the mass spectrometer, the relationship between the assigned and unassigned peaks in the mass spectrum and the size of the database used. For protein identification, advances were made in the analysis of small samples, complex mixtures of proteins and determination of peptide amino acid sequences by tandem MS in which the peptide ion is fragmented in the mass spectrometer into its constituent amino acids. For complex mixtures of proteins, where additional information is required for protein identification, sequence data from fragmentation of peptides by tandem MS are also used to search databases and provide confident assignments of protein identification. More detailed discussion on the development of strategies for protein identification will be given in section 1.3.

***Characterisation***

Characterisation proteomics is the first step in functional protein discovery; investigating the downstream effects of signalling events and protein interactions, for example those driven by post-translational modifications. Post-translational modifications include proteolytic cleavage of part of the sequence, for example removal of a signal peptide or initiator methionine residue, adduction of chemical groups, for example by acetylation, phosphorylation or glycosylation, or the formation of inter- or intra-peptide linkages, for example disulphide bonds. These affect the behaviour of the modified proteins, altering their function or causing them to interact in a different way with other proteins. Various approaches have been applied to characterise these modifications, both targeted and on a larger scale. Of the emerging technologies in this field, top-down proteomics using MS is becoming increasingly popular for characterisation of proteins and proteomes. The determination of the molecular mass of intact proteins by MS is best achieved using high resolution instruments (Jensen *et al.,* 1999) but complex mixtures of proteins can also be resolved (at lower resolution) using bench-top instruments (Hayter *et al.,* 2003). With significant advances in instrumentation, top-down proteomics is currently beginning to offer complementary information to peptide based approaches using a combination of molecular mass determination and fragmentation of the intact protein (Claverol *et al.,* 2003). In particular, this methodology is used to characterise proteins that are post-translationally modified and the interactions between proteins in various biological systems, for example different cell states after drug treatment or genetic manipulation (Cox and Mann, 2007; Siuti and Kelleher, 2007). For analysis of post-translational modifications in a mixture of proteins,

this approach retains connectivity with the parent protein (rather than the loss associated with digesting to peptides), facilitating assignment of modifications to specific proteins. Upon fragmentation, ions are easily reconciled to the parent protein, enabling the location of modification to be detected. With most post-translational modifications resulting in variations in molecular mass and net charge of the protein, isoforms can be separated using 2D gel electrophoresis, and are observed as proteins with similar molecular mass but varying isoelectric point. For example, adduction of acetyl- groups by acetylation of ε-amino groups of lysine residues results in a mass increase of 42Da per acetyl- group, and a decrease in net charge of -1. The decrease in net charge, reflected in the isoelectric point of each isoform will be detected by gel electrophoresis, but this does not provide information as to the specific identification and location of the modification, and moreover does not give an accurate measurement of molecular mass with and without modification. For acetylation, each protein spot on a gel may be a mixture of molecules in which the total number of acetyl- groups is the same, but the sites of acetylation vary. For mono-acetylated proteins, the modification could be at a number of different lysine residues, the specificity of which may be functionally significant (Turner, 2002). To characterise protein isoforms, intact proteins may be analysed by MS directly or by a combination of extraction via passive elution from SDS gels and removal of SDS prior to MS analysis of the intact protein. This confirms molecular mass and predicts the likely nature and number of post-translational modifications based on mass differences between isoforms. Successive rounds of protein fragmentation confirm labile modifications, for example phosphorylation as sequential loss of one or more phosphate ions from the intact protein precursor indicates mono- or di-phosphorylation. These ions may be fragmented further in the mass spectrometer to confirm tentative assignments of multiple modifications and reveal information about the order in which certain modifications have occurred. To ascertain the location of modifications, proteins are digested in-solution and peptide ions are fragmented, revealing sequence differences relating to the modification (Claverol *et al.*, 2003). This combines bottom-up approaches with top-down, as both strategies offer distinct advantages for protein identification and characterisation, and are likely to co-evolve for global proteomic analysis, rather than domination of the field by one or the other (Chait, 2006). As this technology is still developing, there are many challenges that must be overcome if its use is to become widespread. Firstly is the importance of high mass accuracy requiring specialist mass spectrometers, thus limiting their application in many proteomics facilities. Concomitantly is the limited amount of intact mass data currently in proteomics databases, although this is likely to improve substantially over the next few years. On the sample preparation side, top-down

proteomic strategies require a considerably greater amount of starting material to complete multiple analyses. However, this is matched with an increase in the amount of complementary data, for example, characterisation of protein isoforms with similar intact mass values, based on fragmentation patterns in tandem MS. With continued development to instrumentation, and software tools for data processing, a combination of top-down and bottom-up techniques will become high-throughput for the characterisation of post-translational modifications on proteins within cellular pathways (Siuti and Kelleher, 2007).

### *Quantification*

The identity and function of a protein does not provide sufficient information to define the extent of change in protein abundance in response to various stimuli, in different physiological or pathological states. It is imperative that protein abundance can be quantified for a comprehensive understanding of biological systems in addition to the therapeutic benefits for example, the discovery of markers for disease (Sebastiano *et al.,* 2003). Quantification strategies fall into two categories; **relative quantification** in which proteins are quantified in one state relative to a second state, for example a fold change difference in protein abundance between two physiological conditions, or **absolute quantification** in which protein abundance is measured explicitly, for example as nmol per gram of tissue. Early methods for quantification proteomics consisted of quantifying fold changes from proteins separated by gel electrophoresis (for example, Luftig *et al.,* 1974). This allowed the identification of small numbers of proteins associated with the onset and progression of disease, in addition to any related changes in expression. With advancing techniques in sample processing and MS, quantification proteomics is at the forefront of systems biology research and detailed discussion on developments in this field will be given in section 1.5.

## 1.2 Stable Isotopes and their Significance for Proteomics

Stable isotopes are naturally occurring, non-radioactive forms of an element containing a different number of neutrons which give rise to a difference in mass. Stable isotopes of the same element have the same number of protons and consequently, the same chemical properties. For proteomics, the most commonly used are stable isotopes of the elements carbon, nitrogen, oxygen, hydrogen and sulphur. Due to the similar chemical characteristics of stable isotopes of the same element, they are used in place of the more abundant isotope as a diagnostic tool whereby the mass difference can be distinguished using MS.

### 1.2.1 Isotope abundance

Most biologically relevant elements have two or more stable isotopes with the lightest being in greatest abundance. As such, stable isotopes of a particular element giving rise to an increase in mass are often referred to colloquially as 'heavy', and the same expression will be used in this thesis. The natural abundance of stable isotopes most commonly used in proteomics is given in Table 1 (www.isotrace.co.nz), although this is subject to variation (Gannes *et al.,* 1998). Deuterium; [$^{2}$H], the stable isotope of hydrogen is used as an isotopic tracer, for example in 'heavy' water to study metabolic pathways with analysis by MS. For proteomics, [$^{2}$H] is used to label chemicals, for example trideuteroacetate, to differentially label two samples (one with acetate, one with trideuteroacetate) during acetylation of peptides (Ji *et al.,* 2000). [$^{18}$O], the stable isotope of [$^{16}$O] lends itself to climate research in which isotope ratio in layers of ice from different years can provide information, for example the original temperature of precipitation. [$^{18}$O] is also widely used in proteomic research to label peptides during proteolysis for quantitative comparisons of two samples (Yao *et al.,* 2001). Both isotopes of oxygen and hydrogen are used in biochemical research for incorporation into organic molecules to monitor changes under particular conditions, or to facilitate structural determination, for example of proteins, carbohydrates and nucleic acids. Stable isotopes of carbon; [$^{13}$C] and nitrogen; [$^{15}$N] are also commonly used in biochemical research, particularly for metabolic labelling with subsequent metabolic flux analysis (Yang *et al.,* 2005), quantitative proteomic analysis (Wu *et al.,* 2004), and as internal standards in mass spectrometry (Gerber *et al.,* 2003). Detailed discussion on stable isotope labelling strategies for protein quantification is given in section 1.5.

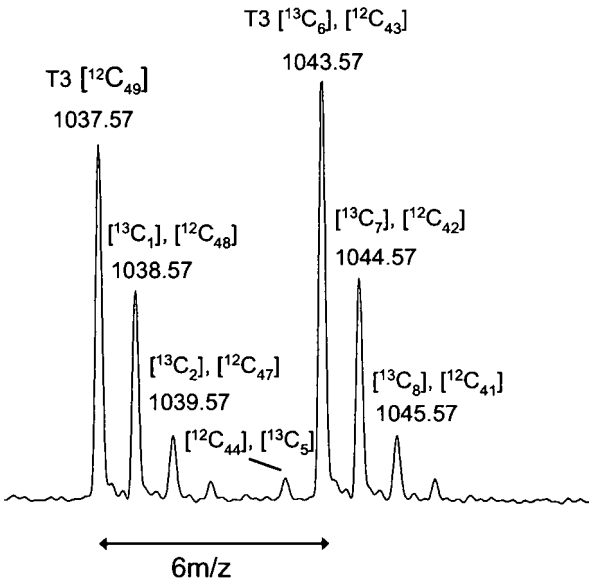
### 1.2.2 The ability of MS to resolve isotopically labelled peptides

The natural abundance of the stable isotope [$^{13}$C], is reflected in mass spectra of all organic compounds, including peptides (Figure 1a). For a peptide containing several carbon atoms, 1.1% of these will be [$^{13}$C] rather than [$^{12}$C]; this corresponds to 1.1% for each carbon atom; for a molecule containing 100 carbon atoms (approximately 10-20 amino acids), there will also be a peak 1D heavier in mass (M+1) and 110% of the height of the main peak (100x1.1). In a mass spectrum of a peptide, this can be use to determine the number of carbon atoms. A 'peptide envelope' also contains a peak corresponding to the peptide containing two (M+2) or more [$^{13}$C] atoms, the intensity of which depends on the number of carbon atoms in the peptide. These isotopic peaks are separated by a constant mass to charge ratio (m/z) which is used to infer the mass of the peptide following ionisation and detection by MS. The charge state of the ion; the number of protons [$H^{+}$] bound to the peptide during ionisation can be

| Element | Stable isotope | Abundance on Earth (%, w/w) |
|---|---|---|
| Hydrogen, [$^{1}$H] | Deuterium, [$^{2}$H] | 0.03 |
| Carbon, [$^{12}$C] | Carbon, [$^{13}$C] | 1.2 |
| Oxygen, [$^{16}$O] | Oxygen, [$^{18}$O] | 0.2 |
| Nitrogen, [$^{14}$N] | Nitrogen, [$^{15}$N] | 0.39 |
| Sulphur, [$^{32}$S] | Sulphur, [$^{33}$S]<br>Sulphur, [$^{34}$S]<br>Sulphur, [$^{36}$S] | 0.78<br>4.55<br>0.02 |

**Table 1. Natural abundance of stable isotopes most commonly used in proteomics**
www.isotracer.co.nz

a.




b.

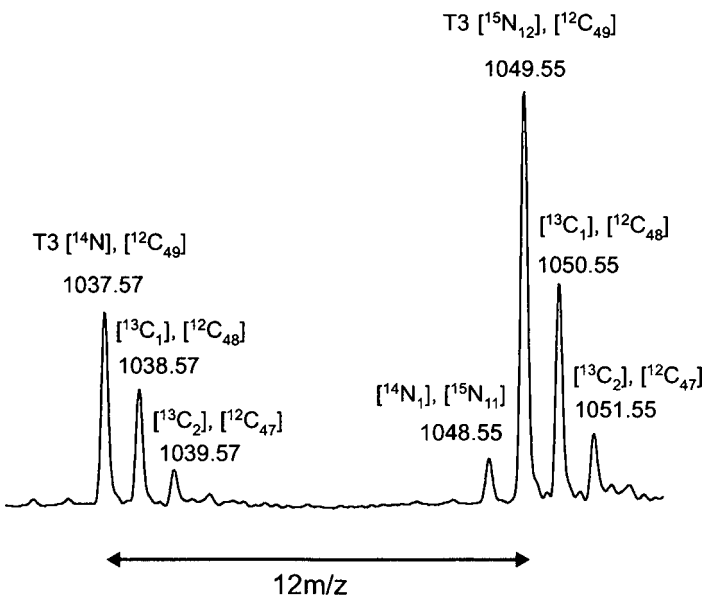



**Figure 1. Stable isotope labelling of a tryptic peptide.**
The tryptic peptide (labelled T3) of sequence GFLIDGYPR $[M+H]^+$ 1037.57m/z, and empirical formula $C_{49}H_{73}N_{12}O_{13}$ labelled with $[^{13}C_6]$-arginine (a) and $[^{15}N]$ (b).

directly inferred by measuring the exact distance between adjacent isotopic components. This principle can also be used to test the efficiency of isotopic modification of peptides and proteins, for example if all [$^{14}$N] atoms are replaced with [$^{15}$N], the abundance of a peak 1Da lighter in mass than the main peptide ion (M-1) resulting from a single [$^{14}$N] atom can be used to report the purity and efficiency of labelling with the heavy isotope (for example, Krijgsveld *et al.*, 2003, Figure 1b).

### 1.2.3 Mass isotopomer distribution analysis

For proteomics, stable isotopes are used in conjunction with MS to monitor and measure changes in abundance at the protein level in biological systems, for example in response to fluctuations in physiological or pathological condition (Hellerstein, 2004). The presence or absence of certain stable isotopes can also contribute extra information to a biological study, for example interpretation of peptide envelopes in mass spectra can be used to report the number of carbon atoms in a peptide and incorporating stable isotope labelled tags that target specific amino acids can be used to count the number of that residue in the peptide in accordance with the mass offset from the analyte. This information can add an extra dimension to protein identification and consequently techniques have been developed to incorporate multiple stable isotopes (for example [$^{13}$C] and [$^{15}$N]) into proteins creating alternative modifications at the peptide level which can be distinguished by mass spectrometry (Snijders *et al.*, 2005). Stable isotopes are used for mass isotopomer distribution analysis by measuring incorporated stable isotope label in the precursor pool and tracing the distribution of the isotope label to determine protein dynamics (Papageorgopoulos *et al.*, 1999). For incorporation into polymeric protein complexes, precursor subunits are labelled and administered combining labelled and unlabelled subunits. Administered precursor subunits are enriched with stable isotopes in different ways, for example replacing $H_2O$ with [$^2H_2$]O in culture media or solvents and the resulting isotopomer profile determined by MS is compared to the expected distribution to establish the isotopic enrichment of the precursor pool (Fanara *et al.*, 2004). For proteins with high molecular weights that cannot be easily resolved using standard MS instrumentation, this analysis can be based at the peptide level using selectively labelled amino acids to determine mass isotopomer distributions. The rate of incorporation of the labelled peptide will represent synthesis of the intact protein. This can provide useful information to understand the dynamics of biological systems, in addition to therapeutic benefits, including the use of stable isotopes to monitor the response of polymeric molecules to certain drug treatments (Fanara *et al.*, 2004).

### 1.2.4 Quantitative proteomics

MS is not inherently quantitative, particularly for peptides, as there are many factors contributing to the loss and generation of ions (discussed in section 1.3.2 and 1.4.2). Consequently, quantitative proteomics often employs stable isotope labelling to compare protein abundance across multiple samples (Cox and Mann, 2007). Stable isotopes have been used in this way for over 30 years when they replaced the use of hazardous radioactive isotopes for quantitative analysis, (for example, Weinkam *et al.*, 1978; Browne, 1986). This can take the form of tags containing stable isotopes which are attached to proteins or peptides through targeting of specific amino acid residues, functional groups or through enzyme catalysed reactions, labelling of whole proteomes through metabolic incorporation of stable isotopes, or using isotopically labelled internal standards (Cañas *et al.*, 2006). These strategies measure changes in protein abundance at the level of mass spectrometric analysis with peptide isoforms separated in mass spectra according to the m/z difference of the stable isotope used (Figure 1). There are several important considerations when designing a stable isotope labelling experiment for quantification of proteins (Julka and Regnier, 2004); firstly is whether all peptides, or selected peptides are tagged. The advantage of a global approach is that all peptides are potential candidates for quantification, however selecting specific amino acids or functional groups can incorporate a significant degree of simplification of the proteome simultaneously, providing that important peptides or proteins are not discounted, thus losing vital information. It is also important to consider the mass window between peptide isoforms, as quantification will become complicated by overlapping isotope envelopes of other peptide species, or from the [$^{13}$C] peaks of the light isotope. Labelling must also be quantitative; if basing abundance on relative signals in MS, it is important that the labelling strategy does not produce a difference in signal that may be incorrectly assigned to the natural abundance of the peptide. It may also be beneficial to multiplex quantification for more than two samples, or more than one analyte, thus quantifying the abundance of several proteins simultaneously. For this, it is important that processing between different samples is consistent, with particular attention to sample processing after labelling for differentially labelled samples. When designing or developing such technologies, it is beneficial to consider how useful the strategy will be for the quantification of proteins in a variety of organisms and using alternative methods of analysis, for example MS platforms.

***Differential labelling strategies***

Stable isotopes of carbon [$^{13}C$], hydrogen [$^{2}H$], nitrogen [$^{15}N$] and oxygen [$^{18}O$] are most commonly used in labelling strategies for proteomics and protein quantification, each offering distinct advantages. Deuterium [$^{2}H$] is commonly used to label many chemicals, for example 'heavy' water, or selected amino acids, as a relatively inexpensive strategy for stable isotope labelling. A drawback of this label is that for use in samples where a peptide separation step is required prior to mass spectral analysis, isoforms containing 'heavy' and 'light' hydrogen do not co-elute when separated by reversed phase; deuterium labelled peptides elute first, as the deuteron is much smaller than the proton, and less hydrophobic (Zhang *et al.,* 2001). Strategies for proteome simplification will be discussed in section 1.4.2. The degree of separation between the two isotopomers depends on the number of deuterium atoms incorporated into the labelled peptide (the greater number of deuterium atoms, the greater the resolution of the two peptides), but for quantification of protein abundance based on MS signal intensity, it is vital that the entire signals from both heavy and light isotopic variants are analysed simultaneously. Consequently, the chromatographic separation of deuterated and non-deuterated peptides could cause considerable error for protein quantification, and is not an appropriate labelling strategy for high throughput protein quantification. In addition, deuterons, are not necessarily metabolically stable, for example after *in-vivo* metabolic incorporation, the α-carbon deuterium may be lost by transamination (Pratt *et al.,* 2002). As an alternative, [$^{13}C$] and [$^{15}N$] are becoming more widely used, for example [$^{13}C$]glucose or [$^{15}N$]$H_4Cl$ as metabolic precursors, although this incorporates a varying number of stable isotope labels into each peptide, which can complicate analyses. Alternatively, amino acids are selectively labelled, for example [$^{13}C_6$]arginine giving a 6Da mass offset from each arginine terminated peptide, or [$^{13}C_6$][$^{15}N_4$]arginine, giving a 10Da mass offset. For this approach, incorporation of one stable isotope, for example [$^{13}C$] is preferred to minimise cost, and to provide a sufficient mass offset (>4Da) from the unlabelled peptide (reviewed by Beynon and Pratt, 2005). [$^{12}C$] and [$^{13}C$] labelled peptides do not resolve during reversed-phase chromatography, even on incorporation of multiple [$^{13}C$] atoms (Zhang and Regnier, 2002). Stable isotopes [$^{18}O$] and [$^{15}N$] also show no chromatographic differences relative to their light isotope; [$^{15}N$] is commonly used as a metabolic labelling strategy for cells in culture that are grown in [$^{15}N$] enriched media and [$^{18}O$] is also used to label water that is the solvent in a proteolytic digestion reaction, incorporating a labelled oxygen atom onto the carboxyl group of a peptide formed during amide bond hydrolysis (reviewed by Julka and Regnier, 2004).

***Internal standards***

Stable isotopes are also used to create internal standards for protein quantification by MS to compensate for variation in sample preparation and analysis. The principle of internal standardisation is employed to establish the relationship between a measured physicochemical response to an analyte and the amount of analyte producing the response. To achieve this, internal standardisation methods have been developed using radioactive isotopes (for example Byrne and Benedik, 1997), although stable isotopes are preferred for biochemical analysis of proteins and peptides using mass spectrometry. This is necessary as signal intensity in MS e.g. MALDI-ToF MS does not give an indication of quantity of analyte due to ionisation and other effects. In many analytical systems, external standardisation is carried out using a series of calibration standards of known concentration. These are chromatographed or analysed separately from the samples and the data are used to convert detected responses for the samples into accurate masses or concentrations. To determine the absolute amount of an analyte protein, internal standards should be chemically identical and show the same behaviour as the analyte but they must also be discriminable, for example by MS. Providing this is the case, a known amount of an internal standard is added to all analyte samples and the response of analyte and standard is compared for absolute quantification, compensating for variations in sample size, preparation and other parameters. This technique has been developed for quantitative proteomic studies in which a standard is added to a biological sample, the chemical and physical properties of which are taken to be representative of the native analyte. For these analyses, the most commonly used surrogates are isotopically modified versions of the analytes (Gerber *et al.,* 2003). Since these peptides are identical apart from the heavy isotope, they have identical ionisation efficiencies in MS, thus relative MS signals between the two can be used as a measure of their absolute abundance. In this way, the use of stable isotopes permits absolute rather than relative protein quantification acquiring unambiguous information in a systems biology context that is comparable across multiple platforms. The use of stable isotopes to achieve absolute quantification of proteins in biological systems and a more detailed description of the evolution and application of these methods is discussed in section 1.5.

## 1.3 Protein Identification

### 1.3.1 Gel electrophoresis

A mixture of proteins can be simply resolved and visualised using sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) in one or two dimensions; a separation technique based on the movement of charged molecules in an electric field. Dissimilar molecules migrate at different rates, separating the components of a mixture. The electrophoretic mobility depends on charge, size, shape and strength of electric field. As such, proteins can be separated according to molecular weight, which may be affected by higher order protein folding, post-translational modifications and other factors. Most proteins carry a net negative charge, depending on the numbers of positively and negatively charged amino acid side chains at a given pH. Gel based electrophoresis methods use a supporting medium (e.g. acrylamide) which provides a cross linked matrix and acts as a filter that proteins must pass through. This increases resistance and minimises the effects of convection currents and diffusion of protein molecules within the buffer solution. To denature protein structure prior to separation, the anionic detergent sodium dodecyl sulphate (SDS) is mixed with the sample ensuring that each protein has a negative charge that is proportional to its mass. For identification proteomics, proteins are separated according to molecular weight only (1D SDS-PAGE) or separated by charge in one dimension and molecular weight in another (2D SDS-PAGE). This affords a considerable degree of protein separation allowing protein expression across different samples to be compared. Separation in two dimensions can also give an indication of some post-translational modifications that introduce variation in molecular weight or net charge of the protein, for example acetylation (discussed in section 1.1.2) and phosphorylation.

### 1.3.2 Mass Spectrometry

Proteins, either in-solution, or after excision from a polyacrylamide gel, are digested into peptides for analysis by mass spectrometry (a detailed discussion of this process is given in section 1.4.1). Peptides are simpler to analyse than proteins as they have lower charge states following ionisation and their mass to charge (m/z) values are compatible with most instrumentation. There are three main processes of MS, ionisation and transfer of ions into the gas phase (from solid or liquid), discrimination of ions (usually by separation or selection) based on their mass to charge ratio (m/z), and detection. To achieve this, there are three main components of a mass spectrometer, the ion source which creates charged particles and

transfers analytes into the gas phase, a mass analyser which separates ions according to their m/z, and the detector which detects the ions of different m/z values in order to generate a mass spectrum.

Two predominant forms of **ionisation** are effective for proteins and peptides, both developed in the late 1980's. **Matrix assisted laser desorption/ ionisation** (**MALDI**; Hillenkamp *et al.*, 1991) ionises peptides from the solid phase, and **electrospray ionisation** (**ESI**; Fenn *et al.*, 1989) creates peptide ions from samples entering the mass spectrometer in solution (Figure 2). The principle of ionisation for both types of ion source is fundamentally the same, as analyte molecules are transferred to the gas phase and ionised. As analytes enter the gas phase, they encounter proton donors [$H^+$] from the matrix (MALDI), or carrier solution (ESI), giving them a positive charge (some applications of mass spectrometry also detect deprotonated/negatively charged ions). MALDI usually adds a single proton [$H^+$] to each peptide, usually on the basic C-terminal residue lysine or arginine for proteins digested with trypsin, thus the entire signal from that peptide is contained within a single entity $[M+H]^+$. Occasionally, MALDI may produce doubly $[M+2H]^{2+}$ or triply $[M+3H]^{3+}$ ions for multiple charge sites, whereas with ESI, several charges may be applied to peptides and proteins, depending on their mass and available protonation sites. For tryptic peptides, ESI usually results in the formation of doubly charged ions, with protons sequestered on the C-terminal basic amino acid arginine or lysine, and on the N-terminus. However, peptides containing a missed cleavage site, for example –ArgPro-, or histidine residues, may be multiply charged in ESI. Both ESI and MALDI ion sources have been coupled to various forms of mass analyser and detection system, thus producing a wide variety of different instruments offering individual analytical strengths. There are three main specifications of a mass spectrometer that deliver different performance from each instrument; sensitivity, mass accuracy and resolution.

**Sensitivity** is dependent a great deal on ionisation, ion transmission and detection. For complex mixtures of peptides, ESI confers greater relative sensitivity than MALDI due to ionisation suppression effects in MALDI, although it is more susceptible to impurities, for example salts used in buffers and solvents (Yang *et al.*, 2007). Both forms of ionisation favour different peptides; MALDI favours basic residues and the side chains of certain amino acids, whereas ESI favours hydrophobic amino acids (Stapels *et al.*, 2004). As such, for protein identification, the two forms of ionisation offer complementary techniques (Stapels *et al.*, 2004). The presence of multiply charged ions may complicate mass spectra when using ESI over

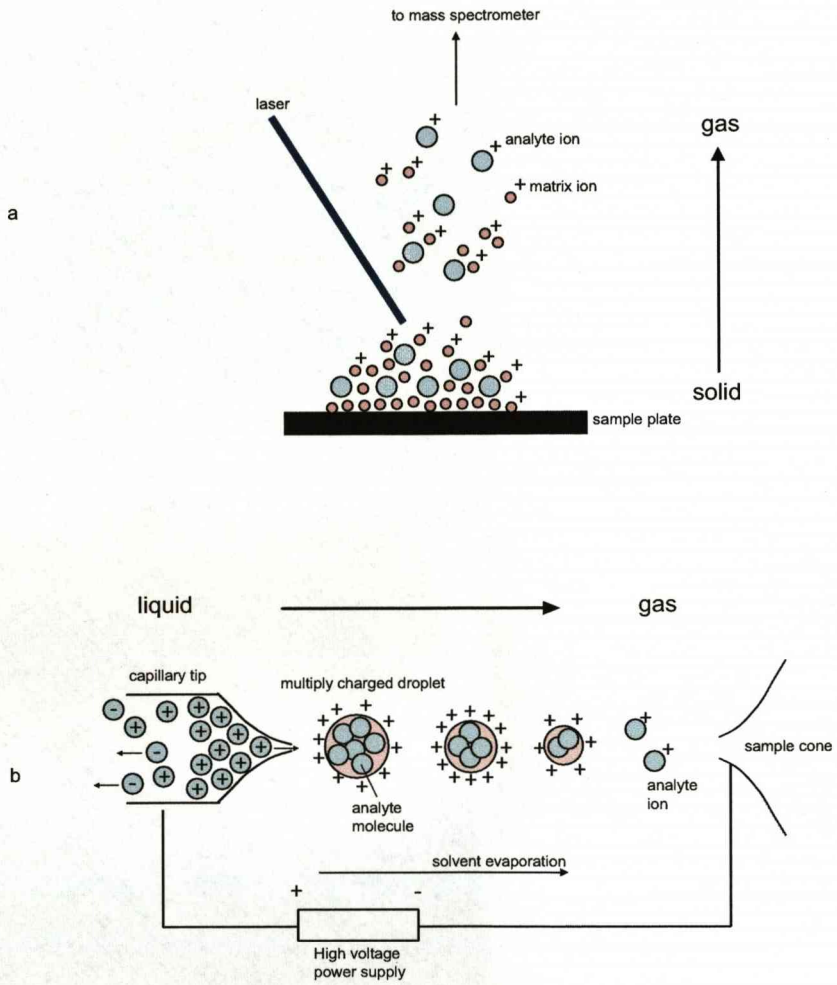
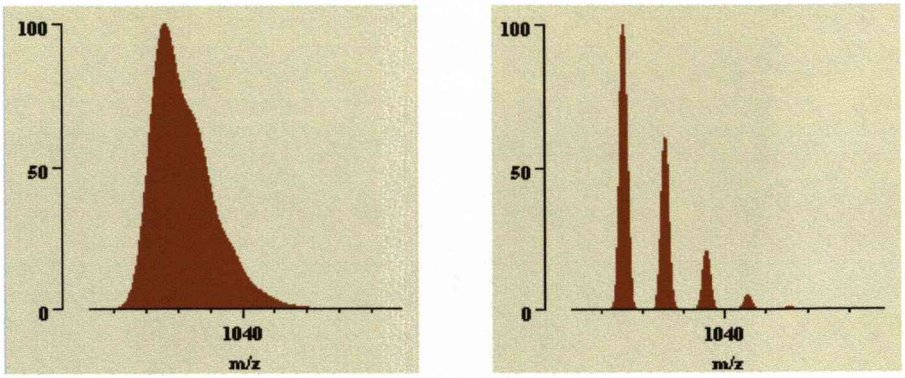



**Figure 2. Ionisation by MALDI and ESI for mass spectrometry.**
For protein identification by mass spectrometry, ionisation is achieved by a) matrix assisted laser desorption/ionisation (MALDI) or b) electrospray ionisation (ESI). MALDI is a solid phase technique in which analyte sample is mixed with an acidic matrix and dried on a target plate. The matrix absorbs energy at the wavelength of an applied laser, causing analyte molecules to be irradiated, vaporised into the gas phase and ionised (protonated; $[M+H]^+$). For ESI, sample containing peptides is sprayed from a high voltage needle into the electrospray source which is maintained at a constant potential difference across the sample cone. Solvent evaporates as a dry gas, for example nitrogen is applied causing the charge density of each droplet to increase. Eventually the charge density reaches a critical level (the 'Rayleigh' limit) and ions are ejected and enter the mass spectrometer.

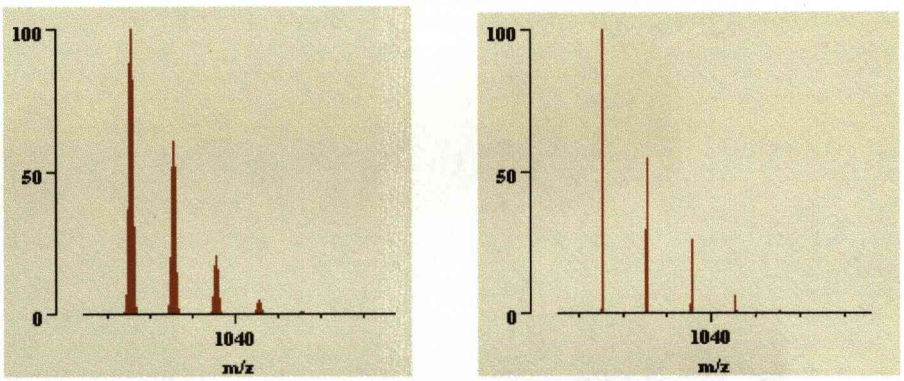
MALDI and the signal from any particular ionised species may be distributed over several charge states. While this may increase analysis time by confirming the charge state of each anayte:internal standard pair, peptide ions in different charge states should behave in the same way, thus not compromising quantification.

**Resolution**; the ability to distinguish between ions of different m/z in a mass spectrum, is largely dependent on the mass analyser and will indirectly confer **mass accuracy**. The capacity of an instrument to distinguish between two signals that are close together is dependent on the mass of the compound and the width of the mass spectral peak measured at different intensities. As a standard measurement of resolution, the full width of the peak is taken at half the maximum height (FWHM) and divided by the mass; the better the resolution, the more accurate the m/z and inferred mass value (Figure 3). To achieve high resolution and mass accuracy appropriate for various applications, modern mass spectrometers use several types of mass analysers; principally time of flight, quadrupoles, ion traps and ion cyclotron resonators (the relative merits and drawbacks in terms of resolution and mass accuracy are reviewed by Balogh, 2004).

**Time of flight** mass analysers (ToF; Stephens, 1946, Figure 4a) employ a flight, or drift tube along which ions travel following acceleration from the source with kinetic energy directly related to their mass and velocity ($KE=\frac{1}{2}mv^2$); smaller ions travel along the flight tube faster than larger ions with less kinetic energy. As all the ions travel the same distance with the same kinetic energy, the time taken can be measured as a direct result of their mass. Resolution is dependent on the distribution of flight times for ions of the same m/z, which is a direct result of slight fluctuations in kinetic energy as ions are released into the flight tube. To increase the resolution of ToF analysers, flight tubes are typically lengthened using a device called a 'reflectron' or 'ion mirror' which is added to the top of the flight tube. This is a focusing device that creates an electrostatic field, deflecting the ions off axis to the detector. The reflectron compensates for small differences in kinetic energy of ions with the same m/z, as those with greater kinetic energy travel further into the reflectron. Consequently, ions of the same m/z reach the detector simultaneously, dramatically improving instrument resolution from around 8,000 to 15,000 (FWHM). The reflectron restricts the mass of the molecules that can be analysed by MS as ions with large m/z values (>5000) cannot be sufficiently deflected by the reflectron, thus decreasing sensitivity and resolution significantly. Resolution can be improved using MALDI and ToF MS by delaying the pulse of ion extraction, focusing the ions immediately




Resolution (FWHM) 1,000 5,000




Resolution (FWHM) 10,000 100,000

**Figure 3. Resolution of peptide mass spectra.**
For mass spectrometry, resolution is a measure of the ability to distinguish between ions of different m/z in a mass spectrum. The capacity of an instrument to distinguish between two signals that are close together is dependent on the mass of the compound and the width of the mass spectral peak measured at different intensities. As a standard measurement of resolution, the full width of the peak is taken at half the maximum height (FWHM) and divided by the mass. To achieve high resolution and mass accuracy appropriate for various applications, modern mass spectrometers use several types of mass analysers with resolution ranging from 4,000 (quadrupoles) to $10^6$ (FT-ICR). Resolution from 1,000 to 100,000 is illustrated for a peptide of sequence GFLIDGYPR, mass 1036.54Da (images created by MSIsotope; http://prospector.ucsf.edu/).

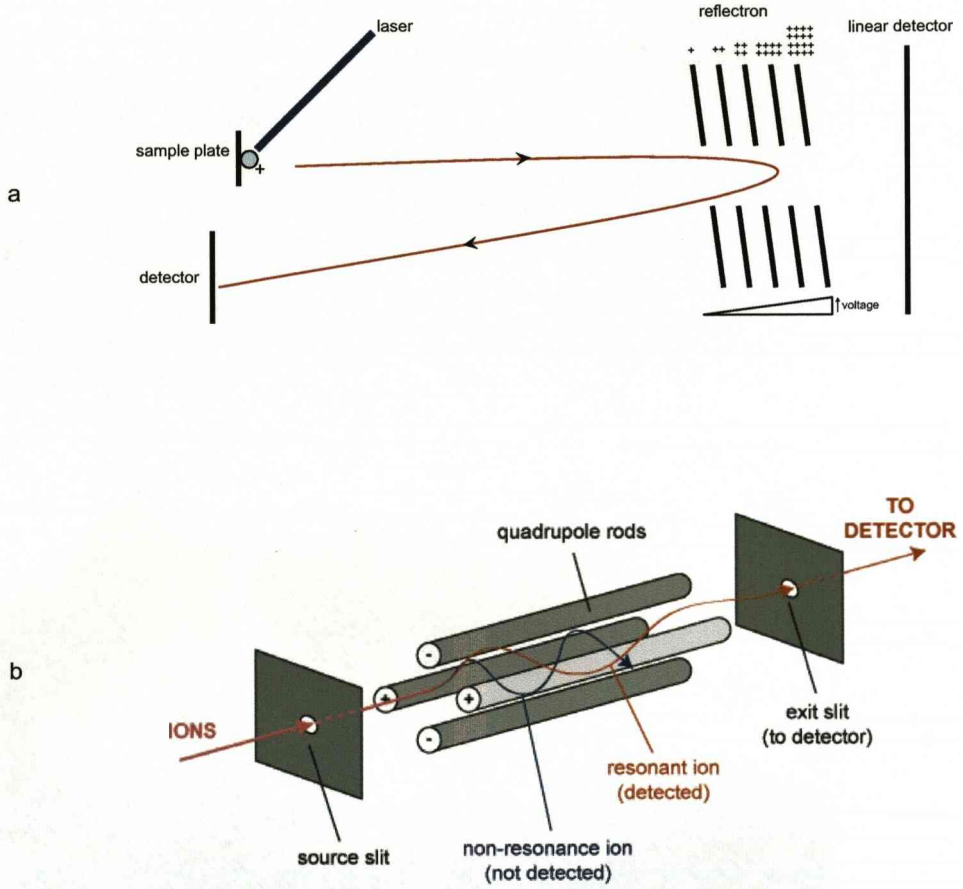



**Figure 4. Time of flight and quadrupole mass analysers.**
a) Time of flight with reflectron; once ionised, the mass to charge ratio (m/z) of ions can be detected in a time of flight (ToF) mass analyser. Ions are accelerated from the source with a kinetic energy dependent on their mass and velocity ($KE=\frac{1}{2}mv^2$). Measurement of the time taken for each ion to reach the detector is indicative of its mass at its particular charge state; lighter ions travel faster with greater kinetic energy and reach the detector first. To improve resolution of the resulting mass spectrum, flight tubes often incorporate reflectrons or ion mirrors that lengthen the path of the travelling ions. Ions with the same m/z and varying kinetic energy are reflected to a different extent as ions with greater kinetic energy travel further into the reflectron, compensating for slight differences in flight time, causing them to reach the detector simultaneously.
b) Ions are separated according to their mass to charge ratio in a quadrupole using electric fields created by four oppositely charged rods. The electric field is created by a fixed direct current and alternating radio frequency across the rods, through which ions of a particular m/z will have stable trajectories, allowing them to pass through to the detector. (Image taken from www.bris.ac.uk).

following ionisation. This generates ions with a significantly smaller kinetic energy distribution so that all ions of the same m/z enter the flight tube at the same time, and with the same kinetic energy. Delayed pulsed extraction is set for the mass range at which optimal resolution is required, with effects significantly diminished above m/z values of 30,000. Resolution of ToF instruments is typically in the range of 10,000 with peaks derived from the same peptide showing clear baseline separation for the carbon-isotopes, and mass accuracy of 200ppm (1000Da±0.2Da). For reflectron ToF instruments, resolution of up to 15,000 can be achieved with mass accuracy of 10ppm±0.01Da).

**Quadrupole** mass analysers (see reference for quadrupole ion trap; Figure 4b) separate ions due to the electric fields created by four parallel rods. Opposite rods are electrically connected in parallel to a radio frequency generator, and direct current. This creates an oscillating electric field allowing only ions with stable trajectories; measured as a function of time and position of the ion from the centre of the rods to pass through and reach the detector. The radio frequency field applied across the quadrupole transmits ions of a specific m/z, allowing only these to pass through and reach the detector. Scanning the RF field allows a broad m/z range (100-4000) with typical resolutions of around 4,000.

The **quadrupole ion trap** (QIT; Paul, 1990, Figure 5a) exists in linear and three-dimensional forms (March, 2000). In contrast to standard quadrupoles, ions are retained inside the trap and will remain so for the time taken to perform standard MS experiments. A three dimensional quadrupole consists of a ring electrode and two hyperbolic endcap electrodes, 'trapping' the ions on a three-dimensional trajectory within a radio frequency quadrupole field. All ions enter the trap and a resonant frequency of dynamic amplitude can be applied to the quadrupole to destabilise successive ion trajectories, thus expelling ions of a selected mass. In these instruments, resolution is inversely related to scanning speed; lower scan speeds increase the density of data points per unit m/z, thus increasing resolution. For high throughput, higher scanning speeds are needed, thus sacrificing superior resolution (typically around 8-10,000, although resolution of 30,000 can be achieved with lower scanning speeds) and sensitivity, as only a limited number of ions can be retained in the trap at one time. A linear ion trap retains ions along the axis of a quadrupole to which a combination of two-dimensional radio frequency voltages are applied to the rods and a direct current is applied to the end lenses at both sides of the trap. This increases the capacity to retain trapped ions, thus increasing sensitivity and dynamic range.

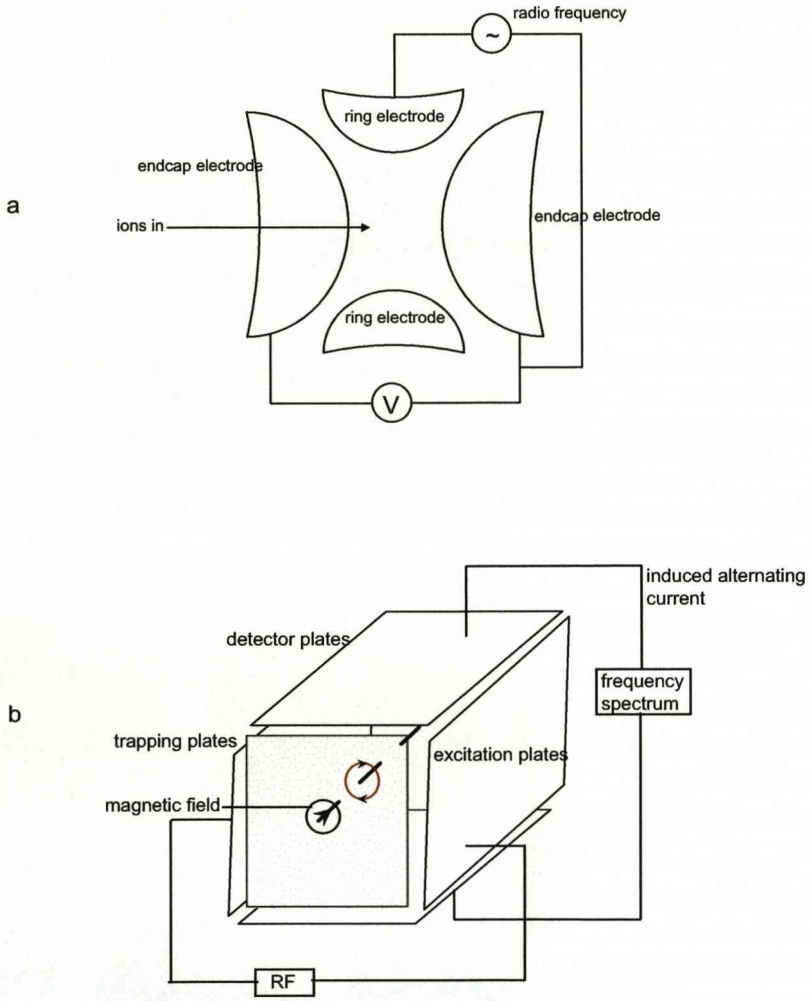



**Figure 5. Quadruople ion trap and Ion cyclotron resonance.**
a) Quadrupole ion trap: Ions are trapped in an oscillating electric field applied by a radio frequency alternating current and a fixed direct current. Specific ions are selectively retained in the trap by applying a resonant frequency to the quadrupole that will successively destabilise ion trajectories causing ions with a particular mass to charge ratio to be ejected from the trap.
b) Ion cyclotron resonance: Ions are trapped in a magnetic field causing them to oscillate with a frequency inversely related to their mass to charge values and directly related to the intensity of the magnetic field. The magnitude of their oscillations is increased as an alternating current is applied, the frequency of which is detected and transformed to produce a mass spectrum.

**Ion cyclotron resonance** (ICR; Hipple *et al.*, 1949, Figure 5b) traps ions in a cell composed of four electrodes in a strong magnetic field. Inside the trap, ions oscillate with a frequency inversely related to their m/z values and directly related to the intensity of the magnetic field. A radio frequency voltage is applied, increasing the magnitude of oscillation of the ions whilst maintaining their frequency. To detect the ion, the frequency is measured with high accuracy via the creation of an image current in the detector with the same frequency as the ion. This achieves high resolution and subsequent mass accuracy as frequency is easier to measure than voltage, which is improved further with a stronger magnetic field. For the simultaneous analysis of ions with different m/z values, the measured frequency of the ions is converted into a mass spectrum by **Fourier transforms** (FT). For FTICR (Marshall *et al.*, 1998) instruments, resolution of up to $10^6$ can be achieved using the highly accurate measure of frequency to produce mass spectra.

Fourier transform is also used to produce mass spectra from the **Orbitrap** mass analyser in which ions are trapped around a central electrode (Hu *et al.*, 2005). The static electric field is created by this and a barrel-like electrode on the same axis. The frequencies of the ion oscillations as they orbit the central electrode are used to measure mass to charge. This by-passes the need for a superconducting magnet but retains very high resolving power (up to 150,000) by measuring frequency, and consequently improves confidence of protein identifications (Scigelova and Makarov, 2006).

### 1.3.3 Peptide mass fingerprinting

Analysis of proteins digested into peptides with a specific protease by MS generates a mass spectrum in which peptide masses provide a diagnostic fingerprint of the parent protein which can be used to search databases of theoretically digested proteins (Pappin *et al.*, 1993). Mass to charge ratios of peptide ions $[M+H]^+$ are used to search databases, for example the protein sequence database SwissProt (www.expasy.org/sprot); a biological database containing known full-length protein sequences, incorporating a high level of manual annotation, including protein function, with a low level of redundancy such that protein products from the same gene are included in the same database entry. For mass spectrometry applications, MS data are commonly searched against the MS database; MSDB (http://csc-fserve.hh.med.ic.ac.uk/msdb.html), a comprehensive, non-identical protein sequence database, particularly employed for MSMS data. The confidence of a peptide match upon an appropriate database search is represented by a probability score, for example using the

MOWSE algorithm. The MOWSE (molecular weight search) fragment database was originally derived from over 50,000 proteins and included the calculated molecular weights of peptides created by a specific proteolytic enzyme (Pappin *et al.,* 1993). A scoring algorithm was also developed in which the frequency distribution of peptide masses obtained experimentally within a given molecular weight range was compared to the number of potential peptides in the same range. This scoring scheme was built into the MOWSE search and the database could be used to identify a protein based on 3-4 peptide masses determined by MS. This scoring algorithm also takes into account the signal intensity relationship between lower and higher mass peptides and their abundance, to facilitate identification. Since the basic model, bioinformatic tools have been developed considerably to incorporate complex parameters into the search of mass spectrometric data to include modifications at the peptide level and to improve confidence of search results. Although the MOWSE scoring system is still used, MS database searches use probability based scoring in MASCOT (Perkins *et al.,* 1999) or SEQUEST (Eng *et al.,* 1994) search engines, both of which yield similar results (Elias *et al.,* 2005). To increase confidence of protein identification, individual properties of proteins and peptides are incorporated into the database search. These include mass range, protease used, taxonomy (which can be broad, for example 'mammalia', or individual species can be searched, depending on the availability of sequence data in the database) and instrument parameters, including peptide mass tolerance (Zhang and Chait, 2000). Search engines can also incorporate expected protease cleavage sites that may have been missed by the enzyme (missed cleavages) and modifications, for example oxidation of methionine residues that may have occurred during sample preparation or post-translational modifications such as deamidation of asparagine residues (www.matrixscience.com). For peptide mass fingerprinting, probability based scoring algorithms provide a quantitative measure of the significance of a match with confidence of protein identification dependent on availability of protein sequence information in the database, spectral quality in terms of signal to noise ratio (the influence of background noise in the spectrum), resolution and mass accuracy, with highly accurate mass data achieved with FT-ICR MS providing more confident protein identifications (Smith *et al.,* 2002) and sample processing. In particular, the success of peptide mass fingerprinting depends on completeness of proteolytic digestion (discussed in section 1.4.1).

### 1.3.4 Tandem MS

Identification of proteins by mass alone (peptide mass fingerprinting) is effective for analysis of single proteins digested into constituent peptides. However, in complex mixtures of proteins,

connectivity to the parent protein is lost upon proteolytic digestion and MS analysis. Consequently, for protein identification, mass contributes limited information and alternative parameters are required, for example structural information. To obtain amino acid sequence information using MS, peptide ions (precursors) are fragmented, often by collision with an inert gas such as argon or helium (collision induced dissociation; CID). This creates a spectrum where each peak (product ion) represents the loss of sequential amino acids, predominantly from the N- or C-terminus of the peptide during fragmentation. Individual fragments ions are then detected, for example using a ToF mass analyser giving a series of fragment (product) ions which can be used to determine the sequence of amino acids in the peptide precursor. Most commonly, fragmentation occurs at the amide bond between amino acids with the resulting product ion spectrum containing a series of b-ions from the amino terminus of the peptide and y-ions from the carboxyl terminus (Figure 6). As such, the amino acid sequence can be determined from the mass difference between adjacent peaks in a b- or y-ion series relating to the loss of sequential amino acids from the precursor. This fragmentation pattern is particularly observed from doubly charged $[M+2H]^{2+}$ ions, for example tryptic peptide ions generated by ESI for which the b-ion series carries the single positive charge from the amino terminus and the y-ions, the positive charge from the C-terminus. Fragment ion spectra often contain other ions, for example fragmentation may occur at the C-C bond giving rise to a-ions (containing the N-terminus) and x-ions (containing the C-terminus), alternatively dissociation can occur at the N-C bond giving rise to c-ions (containing the N-terminus) and z-ions (containing the C-terminus; Johnson *et al.,* 1987). Internal cleavage and side-chain fragmentation may also occur depending on the peptide sequence, ionisation and fragmentation methods. Immonium ions consisting of a single side chain formed through a combination of a- and y-type cleavage can also be used to report on the amino acids that are present in the peptide.

Precursor ions can also be fragmented by electron capture dissociation (ECD; Zubarev *et al.,* 1998) in which a free electron interacts with a multiply protonated molecule, or electron transfer dissociation (ETD; Syka *et al.,* 2004) in which electrons are transferred via collision between the analyte cations (for example multiply charged peptide ions) and reagent anions (for example derived from volatised anthracene). These methods are useful for the analysis of post-translational modifications that are lost during fragmentation by CID due to the transfer of vibrational energy across all covalent bonds (Siuti and Kelleher, 2007). By contrast, electron capture and electron transfer dissociation introduce low energy electrons along the peptide

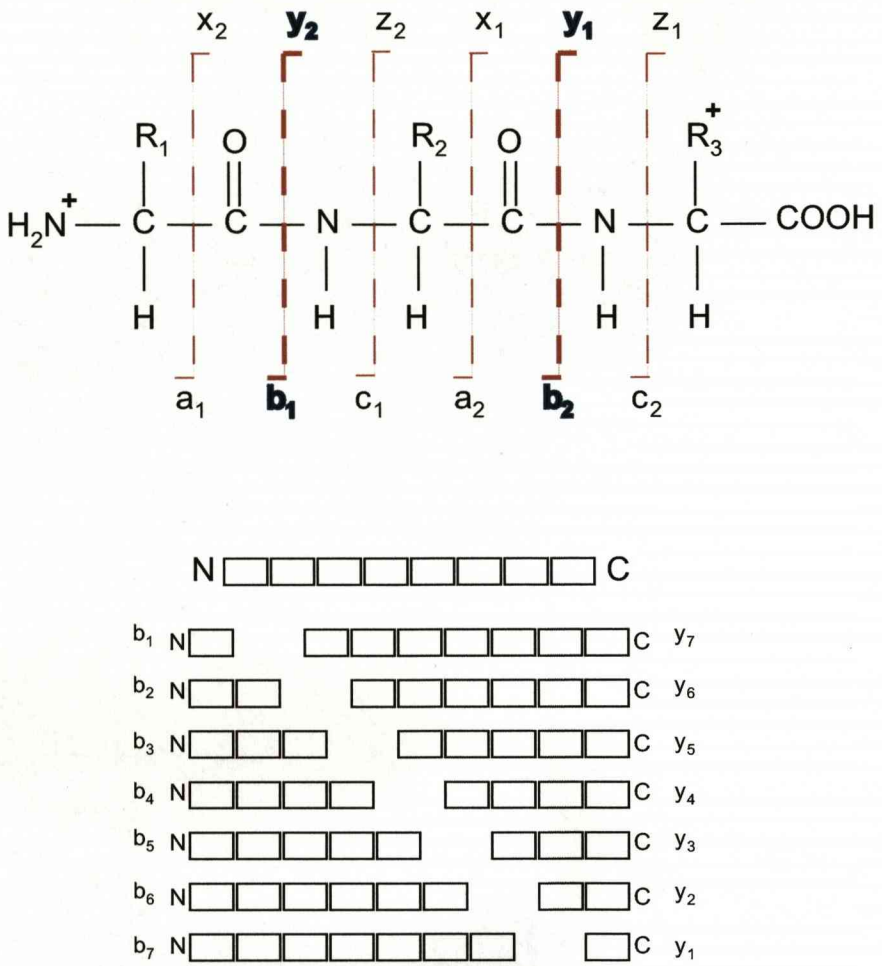



**Figure 6. Fragmentation of a peptide ion.**
For determination of protein primary structure using mass spectrometry, peptide ions are fragmented during tandem mass spectrometry. Fragmentation occurs at the amide bond between amino acids resulting in a product ion spectrum that contains a series of b-ions from the amino terminus of the peptide and y-ions from the carboxyl terminus. The amino acid sequence can be determined from the mass difference between adjacent peaks in a b- or y-ion series relating to the loss of sequential amino acids from the precursor. Fragment ion spectra will often contain other ions, for example fragmentation may occur at the C-C bond giving rise to a-ions (containing the N-terminus) and x-ions (containing the C-terminus), alternatively dissociation can occur at the N-C bond giving rise to c-ions (containing the N-terminus) and z-ions (containing the C-terminus). The diagram above illustrates the b- and y-ions formed from the loss of sequential amino acids from the precursor ion during fragmentation of an 8-amino acid peptide, where each block represents one amino acid residue.

backbone, thus fragmentation is not sequence specific. For MALDI based methods, fragmentation is often achieved by post-source decay where ions undergo fragmentation during time of flight mass analysis (Spengler *et al.*, 1992). Fragment ions are not separated by ToF alone as product and precursor ions have the same velocity and reach the detector simultaneously. However, in reflectron instruments, product ions are separated by the reflectron as they have different kinetic energy values than their precursors. To combine PSD with more rigorous fragmentation using MALDI, instruments have been developed which incorporate a second ToF analyser and a collision cell to fragment precursor ions, for example the AXIMA-ToF$^2$ (Kratos, Manchester, UK). This is useful for combining fragmentation of peptides with the benefits of MALDI, for example its high tolerance of salt contamination, and peptide sequencing is complementary with ESI-MSMS (Noga *et al.*, 2006). However, ESI remains preferable for coupling with quadrupole and ion trap mass analysers for MS and MSMS as it delivers a continuous beam of ions, rather than MALDI which produces short bursts of ions in a vacuum (reviewed by Mann *et al.*, 2001).

Fragmentation in quadrupole ion trap and ICR mass spectrometers is usually achieved by expelling all the ions except the selected precursor. Collision energy is provided through introduction of an inert gas, for example helium, causing the ion to fragment. These instruments are able to perform multiple stages of MS by fragmentation of precursor ions, interrogation of product ions, and subsequent isolation and fragmentation. This technology is compatible with ECD for ICR instruments and ETD for quadrupole ion traps for the analysis of post-translational modifications.

### 1.3.5 MSMS ion searching for protein identification

Tandem MS generated fragment ion data can also be used for peptide and protein identification to search databases using search engines, for example Mascot and SEQUEST. Similar available search tools include X!Tandem (Craig and Beavis, 2004) which matches tandem mass spectra with peptide sequences, and Piums (Samuelsson *et al.*, 2004) which uses a combination of peak extraction and protein scoring for confident protein identification. This can be achieved for single peptides, but is often highly automated (Tabb *et al.*, 2002), with increased confidence in protein identification with multiple peptide matches. Peptide sequencing can also provide information on post-translational modifications to the protein or peptide, for example phosphorylation, and can often locate these modifications within the amino acid sequence (Yates *et al.*, 1995). The success of 'de-novo sequencing' using fragment

ion spectra will depend on the mass accuracy and resolution of the instrument and how well the peptide has been fragmented. However, often small sections of amino acid sequence 'tags' are sufficient to identify a peptide due to the large amount of sequence data available in universal protein sequence databases, for example MSDB. Databases can be searched by comparing fragment ion spectra with theoretical spectral data, thus requiring no prior user interpretation (MSMS ion search), or to combine the molecular mass of the precursor/peptide ion with a short section of the amino acid sequence generated by partial interpretation of the fragment ion spectrum (sequence query; Mann and Wilm, 1994). Furthermore, short sections of amino acid sequences can be searched using the Basic Local Alignment Search Tool (BLAST), which finds regions of local similarity between biological sequences in the protein database, (www.ncbi.nlm.nih.gov) and calculates the significance of protein identification matches.

For large-scale protein identification studies based on MSMS data, there are a wide variety of MS platforms, in addition to database searching software applications/packages that can be used. For systems biology, it is important that these produce reproducible data of the same quantity and quality. For MS, most commonly used instruments Q-ToF and QIT, boast different performance in terms of mass accuracy, resolution and dynamic range. The extent to which data acquired by each instrument is comparable was assessed by Gygi and co-workers (Elias *et al.,* 2005) who discovered that 60% of proteins, with less than half of all unique peptides, were identified by both instruments. This was improved by replicating analysis of the same sample with the same instrument, suggesting that limitations in the number of ions trapped at one time in a QIT, and time for each CID event in Q-ToF have a strong influence on the output. The extent to which protein identification varied across different sample injections, and different instrumentation for analysis prompted development of more stringent validation criteria for accurate and reproducible protein identification across multiple platforms, particularly when proteins are identified by a single peptide (Chamrad and Meyer, 2005). In an attempt to standardise protein identification, bioinformatic tools to handle both MS and tandem MS data along with databases of predicted peptide identification and fragmentation patterns are now publicly available, pooling resources from different laboratories, for example PeptideAtlas (www.peptideatlas.org). In addition, this facilitates the design of protein identification strategies, for example PepSeeker (McLaughlin *et al.,* 2006), a database containing pre-determined fragmentation information from around 200,000 peptides which can be used collectively to predict rules for peptide fragmentation, including relative peak intensity of product ions. These resources combine spectral data and database search information, along with instrument and

software parameters in an attempt to standardise existing protein identification data whilst improving future protein identification strategies.

## 1.4 CHALLENGES IN PEPTIDE BASED PROTEOMICS

### 1.4.1 Proteolysis

For identification and subsequent peptide-based quantification, proteins are most often digested using the serine protease trypsin. Trypsin is a digestive enzyme which is produced in the pancreas as the zymogen, trypsinogen before being transported to the intestine where it is cleaved and hence activated by the enzyme enteropeptidase. For experimental purposes including peptide mass fingerprinting, trypsin is purified, for example from bovine pancreatic trypsinogen, available in high quantities. Cleavage by trypsin is highly specific, taking place at the peptide bond on the C-terminal side of the basic residues arginine and lysine, unless they are followed by a proline residue. It is the positive charge on these particular residues that interacts with the negative charge carried by an aspartic acid residue (Asp 189) in the substrate binding site of trypsin. The abundance of cleavage sites in the majority of proteins; 10% approximate average abundance of arginine and lysine in eukaryotic proteins (Cagney *et al.*, 2003), for trypsin digestion generates peptides of an average length compatible with detection by most mass spectrometers. For example, a 300 amino acid protein would contain 30 cleavage sites on average, generating 31 peptides of 10 amino acids each, with an average mass of 1000Da. Doubly and triply charged tryptic peptides are easily fragmented using CID, with a single proton on both the N- and C-terminus, generating multiple ions (predominantly b- and y-ions), and producing good quality data for database searching and protein identification. For a 24h incubation, used in most standard digestion protocols, trypsin is most effective at 37°C, pH7-9 (Roche Diagnostics, Lewes, UK) under these conditions, trypsin autolysis products will become sufficiently abundant after 24h to complicate mass spectral analysis, in addition to limiting enzyme activity and specificity (Klammer and MacCoss, 2006).

Less commonly used proteases offer alternative cleavage of the peptide bond, which may be advantageous to particular experiments (Table 2). These may also be used in experiments where conditions are unfavourable for trypsin. Endopeptidase Lys-C cleaves at the carboxyl terminus of lysine residues only and is more stable than trypsin. This protease can be used to digest proteins that have been previously solubilised in high salt buffers, including strong solutions of urea. To generate alternative peptide mass fingerprints that may aid in protein

| Enzyme | Specificity | Optimal pH | Mwt (kDa) | # fragments | Av. Length |
|---|---|---|---|---|---|
| Trypsin | C-terminal R/K | 7-9 | 23 | 31 | 10 |
| Asp-N | N-terminal D | 6-8 | 24 | 13 | 23 |
| Gluc-C | C-terminal E | 4-8 | 29 | 19 | 16 |
| Lys-C | C-terminal K | 8-9 | 28 | 15 | 20 |
| Arg-C | C-terminal R | 7-9 | 27 | 16 | 19 |

**Table 2. Commonly used proteases for proteomics.**
Proteases typically used for protein digestion for mass spectrometry analysis are listed with specificity/location of cleavage, optimal pH, mwt, number of fragments expected from a 300 amino acid protein, and the average length (number of amino acids) of each peptide.

identification, proteases Asp-N (cleaves at the amino terminus of aspartic acid) and Gluc-C (cleaves at the carboxyl terminus of glutamic acid, and aspartic acid under some conditions) are also available. Less specific proteases are also used, but these may generate overlapping fragments and are therefore not used as a tool for protein identification by peptide sequencing. In several proteomic experiments, combining peptide identification data achieved using a combination of proteases often achieves the greatest number of identifications, but increases the workflow considerably (Biringer *et al.*, 2006). Differences in amino acid sequence coverage from different proteases were attributed to the availability of cleavage sites within the three dimensional structure of the protein rather than the enzyme efficiency, although alternative proteases to trypsin are more likely to generate larger peptides, beyond the mass range of popular instrumentation that are consequently not detected (Table 2).

***Factors influencing proteolysis and their consequences***

For confident protein identification, efficient digestion to generate all possible peptides for analysis by MS is desired. However, this is not always possible, particularly for a complex mixture of proteins with varying three dimensional structure and availability of cleavage sites. To compensate, missed cleavage peptides retaining one or more hydrolysable bonds can be included in database searches for protein identification. Sample processing conditions may also be adjusted to improve the efficiency of a protease where possible, as although only a small amount of each peptide must be present in order to be detected, a stronger signal is likely to ensue upon MS if a greater proportion of peptide is present. In addition, there are instances in which complete proteolytic cleavage of peptides is essential, for example in quantitative proteomics using stable isotope labelled internal standards (introduced in section 1.2.3). For this strategy to provide reliable quantification, the response of the peptide selected, in relation to the chemically identical, stable isotope labelled internal standard, must be representative of the parent protein, and consequently must be completely cleaved by the protease. To assess the extent of proteolytic cleavage of analyte proteins, influential factors must be considered. These include sample preparation conditions, for example protease to substrate ratio, nature of buffers and solvents, temperature and the time allowed for digestion to take place, in addition to substrate composition; primary structure, higher order structure and protein folding. Other limiting factors include, non specific proteolysis *in vivo* or during sample processing and autolytic cleavage of the protease, limiting its efficiency within the analyte system.

The majority of sample preparation factors are relatively easy to control, for example proteins digested following excision from gels, or in-solution with trypsin are often reduced using thiols such as dithiothreitol (DTT) (Lundell and Schreitmüller, 1997) or by heating. This reduces disulphide bonds that form between cysteine residues, allowing trypsin access to a greater number of potential cleavage sites within the protein. In order to prevent disulphide bonds re-forming, cysteine residues are alkylated, using, for example iodoacetic acid which causes all cysteines to become carboxymethylated. Addition of several denaturing agents, for example 0.1M urea or 0.1% (w/v) SDS can improve the activity of trypsin up to 180% (from 100%), by allowing the protein to adopt a less rigid tertiary structure to facilitate access to cleavage sites (Roche Diagnostics, Lewes, UK). While this may improve digestion efficiency, these reagents are often incompatible with downstream analysis, particularly in liquid chromatography and ESI-MS. Chaotropes for example urea can result in peptide adducts through the formation of isocyanate, limiting use of the data acquired. Denaturing agents used to solubilise proteins prior to proteolytic digestion for example urea and guanidine hydrochloride are often used at concentrations up to 8M which will significantly reduce activity of trypsin; just 1M GHCl reduces trypsin activity to zero (Harris, 1956; Hill *et al.,* 1958; Roche Diagnostics, Lewes, UK). These agents must therefore be removed prior to trypsin digestion. Improved proteolysis can be achieved by heating the protein, although this can result in sample loss through protein precipitation. Digestion can also be enhanced by increasing the concentration of enzyme but this can become expensive if used routinely. Addition of organic solvents including methanol and acetonitrile (up to 80% v/v) can be used to aid solubility and unfolding of proteins (Strader *et al.,* 2006), increasing trypsin activity significantly. Alternative methods involving precipitation of protein with trichloroacetic acid (TCA) or acetone prior to resolubilisation and trypsin digestion have also been effective at denaturing and desalting proteins for more efficient proteolysis (Kim *et al.,* 2006). Proteolysis yields a greater number of peptide and protein identifications when trypsin has been immobilized on a microcapillary column through which the protein mixture to be digested is passed (Klammer and MacCoss, 2006). This provides an automated strategy combining sample cleanup, preconcentration and protein digestion with increased efficiency and stability of enzyme. Immobilised trypsin can be applied to very low concentrations of proteins; however this is most likely due to an increased enzyme to substrate ratio as the protein passes over the beads. Although these methods to improve proteolytic digestion are effective to a varying degree, it is important to realise that incomplete digestion is tolerable for protein identification, providing sufficient peptides in sufficient abundance are detected by MS.

For individual proteins, especially in their native state, certain regions are more susceptible to proteolytic cleavage by trypsin than others (Halsey & Harrington, 1973; Ellison *et al.,* 1995). despite the majority of lysine and arginine residues on the surface of a protein. Primary structure may be influential, with altering polarity and hydrophobicity surrounding the cleavage site having a significant impact on the ability for trypsin to cleave (Monigatti and Berndt, 2005). Primary sequence in particular has three main influences on trypsin cleavage; the effect of proline following an arginine or lysine residue, the effect of the arginine or lysine residue itself and that of negatively charged amino acids such as glutamic acid and aspartic acid immediately surrounding the cleavage site (Siepen *et al.,* 2007). Enzyme specificity is determined by the efficiency of binding to the substrate and catalysis of the reaction (Schechter and Berger, 1967) and this is determined by the length of the peptide, and the specific amino acids that come into contact with the active site (Figure 7). Consequently, enzyme cleavage may be dependent on a larger proportion of the peptide chain other than the two residues immediately flanking the cleavage site. In particular, successful cleavage is dependent on the nature of the substrate binding with the active site of the enzyme, such that amino acids occupying alternate subsites around the active site have the greatest influence on proteolysis kinetics (Schechter and Berger, 1967). However, under denaturing conditions, trypsin will cleave at nearly every lysine-X or arginine-X bond (Hubbard *et al.,* 1998), thus higher order protein structure greatly influences the rate of trypsin digestion. The extent to which trypsin digestion will occur in a complex mixture of proteins is less well known although it is believed that incomplete digestion is more likely for mixtures of native proteins, due to inadequate denaturation, enzyme concentration, incubation time, or due to inhibition by contaminant molecules that compete for protease activity (Klammer and MacCoss, 2006).

### 1.4.2 Sample complexity and dynamic range

The discrepancy between the number of theoretical peptides produced from a tryptic digest of a proteome and that detected by MS creates a challenge for proteomics. It is feasible that a tryptic digest of a proteome could contain $2.5x10^6$ peptides given that an average protein is digested to 20-50 tryptic peptides and a proteome could contain 20-50,000 proteins including post-translational modifications. For identification proteomics, highly abundant proteins dominate all forms of analysis and those in low abundance are often not detected. The yeast proteome contains around 5,000 proteins ranging in abundance from 50 molecules per cell to $10^6$ (Ghaemmaghami *et al.,* 2003), and the *E.coli* proteome contains around 4,000 proteins with dynamic range of abundance detected from 30 to 90,000 molecules per cell (Lu *et al.,* 2006). In

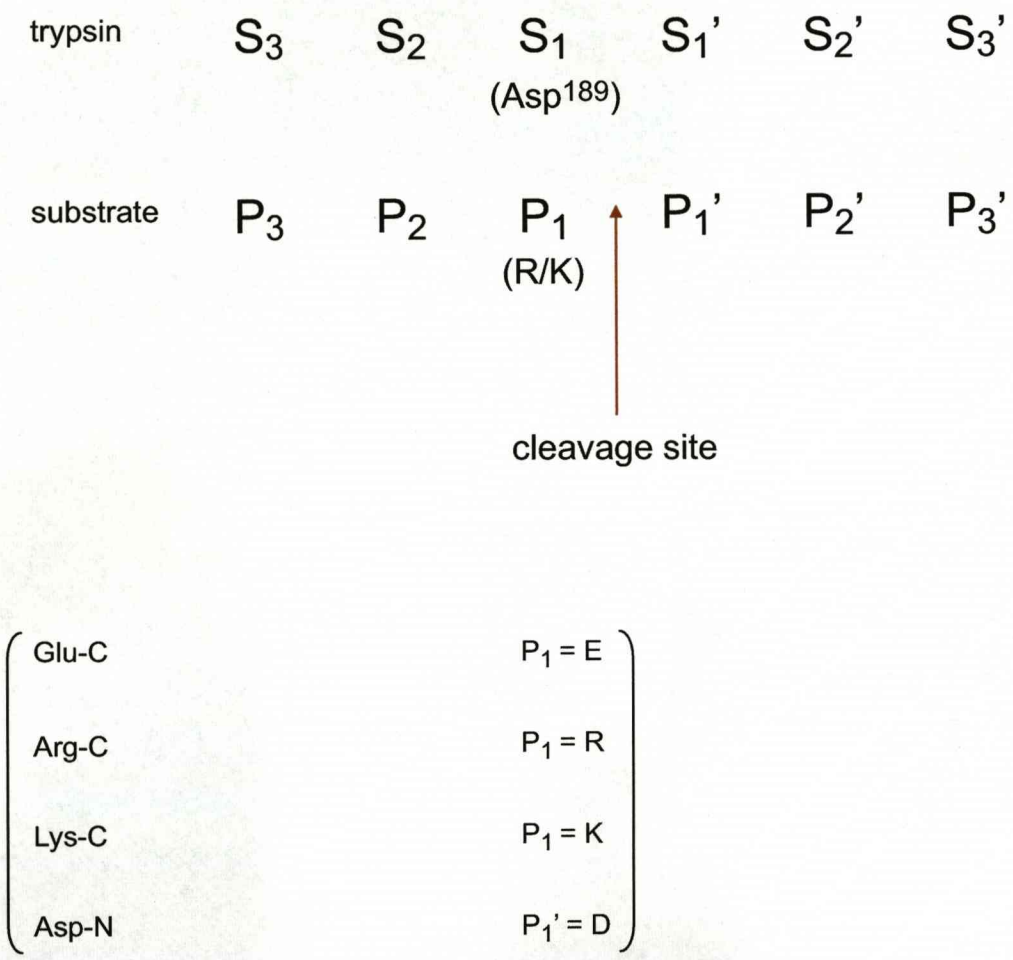



**Figure 7. Nomenclature of proteolysis according to Schechter and Berger (1967).**
Amino acid residues immediately associated with proteolysis are represented; $P_1$-$P_3$ N-terminal to the cleavage site, $P_1$'-$P_3$' C-terminal. Associated subsites on the protease are similarly represented. For the serine protease trypsin which cleaves C-terminal to arginine and lysine residues (except in the presence of proline at $P_1$'), proteolysis is initiated by the formation of a salt bridge between $Asp^{189}$ situated in the pocket of $S_1$ and the basic side chain of arg/lys. Amino acid residues in $P_1$ position for other commonly used serine proteases, Glu-C, Arg-C and Lys-C in addition to the metalloprotease, Asp-N are indicated.

human plasma, the difference in protein abundance covers ten orders of magnitude (Anderson, 2002) between the most abundant protein, albumin and the least abundant measurable proteins. To analyse samples of this complexity, several strategies have been developed. These include multidimensional separation of proteins and peptides prior to MS analysis, the removal of high abundance proteins, and enrichment of low abundance proteins. These strategies can be used to achieve protein profiling directly from complex mixtures without the need for gel electrophoresis, and to reduce sample complexity whilst maintaining the integrity of the original sample.

***Separation of peptides and proteins***

Both peptides and proteins can be separated in-solution by liquid chromatography prior to detection or further processing. Complex mixtures can be separated according to size, charge and hydrophobicity. These separations are predominantly column based, using a carrier solvent compatible with the column chemistry used. The most commonly used separation for peptides prior to MS analysis is reversed phase high performance liquid chromatography (RP-HPLC) where peptides are separated in-solution on the basis of their hydrophobicity. Samples are loaded onto a precolumn or 'trap' at high flow rate where they are concentrated and washed to remove contaminants, including salts. The flow through from this process is diverted to waste and the sample is subsequently loaded onto the analytical column. Both the precolumn and analytical column are packed with silica based beads with surface bound long n-alkyl groups for example n-octadecyl ($C_{18}$), covalently bound. Peptides bind to the matrix and are eluted using a gradient of an organic solution, for example acetonitrile with the most hydrophobic peptides eluting at the end of the gradient in a high concentration of organic solvent. To increase separation efficiency and peak capacity for liquid chromatography, longer columns were developed. This increases the number of protein identifications, but requires very high pressures and is not always appropriate for routine use (Hu *et al.*, 2007). In an alternative approach, polymer based monolithic columns are used for the separation of complex mixtures of proteins or peptides. These operate at low pressure and can be applied to samples with a wide range of pH values. However, the polymer materials may not be stable in some organic solvents and may undergo shrinking or swelling. To combine these methods for more efficient separation of proteins and peptides, two or more chromatography steps are often employed (multi-dimensional protein identification technology; MudPIT). This separates proteins or peptides based on two or more different chemical properties, for example reversed phase separation of peptides may be coupled with prior ion exchange separation of proteins on the

basis of net charge using a salt gradient (Washburn *et al.,* 2001). This increases the time for analysis of each sample, but allows high throughput identification of many more proteins in complex mixtures by MS and has been used for analysis of large molecular complexes (Link *et al.,* 1999). MudPIT is also applied to separations by size exclusion, for example to remove high molecular weight proteins or peptides, and increase the sensitivity of low molecular weight detection, affinity chromatography, for example to selectively recover phosphopeptides or hydrophilic interaction. When used, reversed phase chromatography provides the final dimension of separation due to solvent compatibility with ESI-MS. SDS-PAGE separation can also be used as a further dimension of separation prior to liquid chromatography, or MudPIT analyses. This is typically achieved by slicing the entire gel lane into uniformly sized slices, digesting each one with trypsin and separating by liquid chromatography prior to MS ("GeLC-MS"). This gives an extra dimension of information for protein identification as the approximate molecular weight of proteins in each gel slice is known from their position on the gel (Lasonder, 2002, Steen and Mann, 2004). Peptides can also be separated by isoelectric focusing in-solution using pH gradient strips (Heller *et al.,* 2005), or using a pH gradient in specficially designed media for free-flow electrophoresis (Nissum *et al.,* 2007). Separated peptides are removed in-solution, following separation according to charge, and applied directly to HPLC-MSMS. This technique compared favourably with GeLC-MSMS for identification of proteins from embryonic stem cells, requiring significantly less sample preparation and analysis time (Graumann *et al.,* 2007).

Reversed phased high performance liquid chromatography (RP-HPLC) has been coupled with electrospray ionisation MS (ESI-MS/MSMS) as an effective online method of combining the two techniques of separation prior to fragmentation and mass spectral analysis. However, liquid chromatography now provides a simple and effective front end to a variety of mass spectral detection systems in order to concentrate and clean up a variety of protein and peptide mixtures prior to analysis. The possibility of coupling liquid chromatography to the solid phase MS technique of matrix assisted laser desorption/ionisation time of flight MS (MALDI-ToF MS) has been tested (Mirgorodskaya *et al.,* 2005) as an off-line method to improve the scope of detection by MALDI MS. There are several reasons for combining these techniques, mainly that of user discretion, in that interesting fractions can be re-visited a number of times whilst acquiring data as the peptides are fixed in the solid phase on the plate. Using MALDI-ToF MS to analyse peptides also eliminates the presence of multiply charged species, which often complicate electrospray spectra, particularly in highly complex samples. As chromatographic

behaviour of a peptide is dependent on the amino acid composition and distribution of amino acids along the peptide chain, sequence specific factors can be used to predict retention times during RP-HPLC. This is advantageous for LC-MALDI data, as when combined with information derived from chromatographic retention times to differentiate between peptides, this increases the confidence of protein identification by peptide mass fingerprinting significantly (Krokhin *et al.,* 2004). Off-line fractionation also offers advantages for ESI, especially for MudPIT analyses as separation platforms that are not compatible with MS, or each other can be used. In addition, fraction collection is not limited and certain fractions can be preferentially selected and re-visited for analysis (Hu *et al.,* 2007).

For identification of proteins in complex mixtures, various techniques for protein and peptide separation prior to MS analysis are described. The use of one system over the others is driven by availability of resources and compatibility of the samples to be analysed within the detection system. However, protein identification across different proteomics workflows gives different results, as discussed in section 1.3.5. Comparing protein identification using different approaches for protein and peptide separation prior to mass spectral detection is no different; for example, identification of more than 1000 proteins using three different methods of protein/ peptide separation and analysis, with a comparison of previously published literature, resulted in only 46 proteins identified using all four methods, and 195 proteins located in more than one dataset (Anderson *et al.,* 2004). The reasons for this are not clear, but possibilities may include different methods exposing different subsets of proteins, differences in bioinformatics tools incorporating different degrees of error in identification, or individual sample differences for different experimental strategies.

***Protein enrichment and normalisation***

For identification proteomics, there is a wealth of information suppressed beneath the highly abundant proteins that dominate all forms of analysis. Several attempts have been made to deplete abundant species, for example the most abundant proteins albumin and IgG in plasma, resulting in an increased number of protein identifications following LC-MSMS analysis of peptide mixtures digested from depleted protein samples (Plavina *et al.,* 2007). However, these are often of minimal use in enriching trace proteins, and are not selective in removal of specific proteins (Zolotarjova *et al.,* 2005). Alternative methods, including prefractionation and enrichment of proteins carrying certain post-translational modifications (Zhang *et al.,* 2007) have been developed, but accessing an entire proteome still presents an enormous challenge

for identification and discovery proteomics. In an attempt to address this challenge, an approach to reduce the dynamic range of a complex mixture of proteins has been developed whereby high abundance proteins are diminished and removed from the sample and low abundance or trace proteins are enriched simultaneously (Thulasiraman *et al.,* 2005, Righetti *et al.,* 2006). This methodology uses a library of peptide ligands covalently attached to the surface of spherical porous beads, of sufficient heterogeneity to probe different physicochemical properties of a protein. A typical library consists of linear hexapeptides based on the 20 naturally occurring amino acids, which in theory creates $20^6$ (64 million) different ligands. Only a small subset of ligands would be able to bind specific proteins. In principle, abundant proteins will quickly saturate all of their available binding sites and most will remain unbound whereas low abundant proteins will not saturate all of the high affinity ligands and the majority will bind to the surface of the beads. Exposure to saturating amounts of all proteins would effectively normalise the population of proteins, assuming equal capacity for every protein (Guerrier *et al.,* 2006). This achieves enrichment of trace proteins relative to proteins present in higher abundance. In doing so, the sample cannot be used for quantitative analysis of protein abundance but it is a useful tool for protein identification and has been applied to analysis of human urine (Castagna *et al.,* 2005) and serum (Guerrier *et al.,* 2006), as well as cell and tissue lysates (Righetti and Boschetti, 2007). The number of proteins captured using peptide library beads compares favorably with affinity depletion and liquid chromatography separation technologies with the advantage of fewer fractions, shorter analysis time and significant enrichment of low abundance proteins (Sennels *et al.,* 2007).

***Ionisation efficiency***

Upon proteolytic cleavage of a proteome, generation of tryptic peptides in the region of $2.5x10^6$ would undoubtedly result in a complex mass spectrum, but only a small proportion of the expected peptides would be present. Providing the sample is present in sufficient concentration, and that tryptic digestion has been driven to completion, the predominant reason for this is non-uniform ionisation of all peptides by MALDI or ESI, and the length of time the instrument has to detect each ion as it reaches the detector ('dwell time') in ESI-MS. Previously discussed strategies to reduce sample complexity will also reduce the effects of ion suppression as the presence of a few abundant proteins will impair ionisation of those in lower abundance, particularly in MALDI-ToF MS (Yang *et al.,* 2007). In addition, the signal from peptide ions is often suppressed following MALDI by the matrix itself but this effect can be reduced with alternative sample preparation, for example mixing analyte and matrix in solid

form, rather than applying liquid matrix and allowing co-crystallisation as spots dry (Wang and Fitzgerald, 2001). Ion signal from low abundance species is often suppressed by that from peptides in higher abundance but even then, certain peptides are known to ionise more efficiently than others, with MALDI favouring basic residues and the side chains of certain amino acids, whereas ESI favours hydrophobic amino acids (Stapels *et al.,* 2004). Consequently, this is dependent on physicochemical properties of individual amino acids, for example size, hydrophobicity and charged side chains (Kratzer *et al.,* 1998). This also highlights the propensity for peptides containing certain amino acids to ionise, being dependent on specific amino acid side chains to a greater extent than overall physicochemical properties (Baumgart *et al.,* 2004). In particular for tryptic peptides, MALDI mass spectra are dominated (up to 94%) by arginine, rather than lysine terminated peptides (Krause *et al.,* 1999). This is attributed to the basicity of the guanidino side chain of arginine, aiding ionisation by MALDI. For peptide mass fingerprinting, and improved peptide detection, signal intensity of lysine terminated peptides can be enhanced by conversion of the ε-amine into a guanidino group with the addition of *O*-methylisourea, forming homoarginine (Kimmel, 1967). This method is used to increase coverage of proteins when identified by peptide mass fingerprinting (Hale *et al.,* 2000). This also adds an extra dimension of information when comparing modified to un-modified mass spectra; mass shifted peptides with an increase in signal intensity, or indeed previously unidentified peaks will be predicted to contain lysine residues.

## 1.5 Quantification Proteomics

### 1.5.1 The significance of quantification for proteomics

It is now possible to move forward from identification and characterisation proteomics, to quantification proteomics in which the protein target is known and can be quantified. Quantification proteomics employs techniques to measure relative quantification, and more recently absolute quantification in terms of absolute numbers in a given tissue. This will provide much more comprehensive information about individual proteins in biological systems, rather than defining the concentrations of proteins relative to a second cellular state. The ability to measure minor fluctuations in protein concentrations under different physiological conditions and across multiple platforms brings proteomics into the context of 'systems biology'.

Due to the variable efficiency of ionisation, and detection, MS of peptides is not inherently quantitative, thus signal intensity in MS does not serve as a useful predictor of absolute

abundance, but can be used in combination with internal standardisation using appropriately selected, stable isotope labelled internal standards. In combination with a variety of stable isotope labelling strategies, proteome simplification and MS, multiple proteins can be quantified simultaneously. To achieve absolute quantification based on these precepts, internal standards used must be chemically synthesised, identical replicas of analyte peptides to be quantified containing appropriately selected stable isotopes for discrimination between the two in MS. The evolution of strategies for protein quantification, with and without the use of stable isotope labelling, both relative and absolute is discussed in detail below.

### 1.5.2 Relative quantification

***Label free relative quantification using gel electrophoresis***

Quantitative proteomics using gel-based methods consists of applying algorithms to gel electrophoresis analysis to compare cell populations of different states by detecting spots exhibiting higher or lower stain intensity, reflecting differential protein expression. SDS-PAGE gels stained, for example with Coomassie Blue are scanned and software programmes are used to detect spots and determine semi-quantitative differences in proteins between different samples (e.g. TotalLab, Ceredigion, UK). This is based upon the intensity of the stain, relating to the amount of bound dye, and hence the amount of protein. Such densitometric analysis can be optimised for different instrumentation and stains used (e.g. Vincent *et al.,* 1997). For the comparison of 2D SDS-PAGE gels, spots must be matched on different gels as only one sample can be run on a single gel. To improve this, proteins from two different samples may be labelled with fluorescent dyes (for example, Cy5 and Cy3) prior to simultaneous separation by 2D SDS-PAGE using the technique of difference in-gel electrophoresis (DIGE; Unlu *et al.,* 1997, Viswanathan *et al.,* 2006). To measure differences in protein abundance, the gel is scanned with the excitation wavelength of each dye sequentially, causing fluorescence. This approach, while enabling simultaneous detection of multiple proteins in different samples and eliminating gel to gel variation for analysis of two samples, is time consuming and is most commonly used as a qualitative representation of the two samples. It is also limited to detection of high abundance proteins in a complex mixture.

***Label free relative quantification using western blotting***

Protein abundance can also be measured using western blotting in which proteins are detected using antibodies specific to the target protein. For this analysis, proteins are separated using gel electrophoresis and transferred to a nitrocellulose membrane where they are probed with

specific antibodies. Proteins are probed with a primary antibody raised to the protein of interest and then with a secondary antibody which binds to the primary antibody and can be labelled with a specific probe, resulting in a coloured product to aid detection. Bound probes are detected upon exposure to the specific chemical reagent to produce a signal where the targeted protein is located on the original SDS-PAGE gel. To measure the abundance of this protein, markers of a known amount of a protein significantly separated in molecular weight from the analyte are also detected and the intensity or size of the protein band is compared. For more sensitive detection, chemiluminescence is used whereby the western blot is incubated with a substrate that will cause luminescence when exposed to the secondary antibody. The light signal is detected using photographic film, producing an image which can be analysed using densitometry to measure protein abundance.

***Label free relative quantification using mass spectrometry***

For comparative proteomics, methods that achieve relative quantification without stable isotope labelling or chemical modification have been designed. Protein identification scores (section 1.3.3) do not provide an accurate prediction of protein abundance, as these are largely based on the number of matched sequence fragments and do not incorporate ion signal intensity which is affected by the propensity of each peptide to ionise (Ong and Mann, 2005). Consequently, label free quantification methods use statistical differences in ion signal intensities or chromatographic peak areas as a direct measure of ion signal intensity (number of ions detected by the mass spectrometer at a given time) of a particular ion (Higgs *et al.,* 2005) between several HPLC-MS analyses of the same samples to predict the relative abundance of proteins. This is based on the linear relationship between the signal and concentration of analyte in the chromatographic eluent (Lubec and Afjehi-Sadat, 2007). For complex samples, accuracy is limited by ion suppression effects, but this may be aided by high resolution MS. Extracted ion chromatograms are different for different peptides, but from the most intense peptide ions, they can be averaged to give a quantitative measure of the abundance of a particular protein. As two or more samples are analysed separately, there is no control for variation in sample work-up, thus accuracy is not as high as most stable isotope labelling approaches (Cox and Mann, 2007). However, label free quantification produces abundance data akin to that achieved using 2D SDS-PAGE and densitometry, thus is effective for proteomes that are not analysed easily under these conditions, for example hydrophobic membrane proteins, and for low abundance proteins that are not detected by gel electrophoresis. Likewise there are many peptides that will not be detected by MS as they lie

outside the mass range of the instrument used, or those that do not chromatograph well, for example those that are highly hydrophobic and may remain bound to a reversed phase column. Problems of ionisation can be avoided by comparing an extracted ion chromatogram of the same peptide in two or more samples to give an idea of the relative abundance of a particular protein, as the peptides should behave in the same way under the same experimental conditions (Wang *et al.,* 2006). A significant problem encountered when using the relative intensity of the same peptide between samples for quantification is that of extremely complex chromatograms and resulting mass spectra. In an attempt to increase the throughput of this method, differential MS was developed (dMS; Wiener *et al.,* 2004) as an automated alternative. dMS uses multiple LC-MS runs to select peptides that differ significantly between samples for MSMS. Such an approach targets only differential expression of particular proteins between two samples, thus eliminating uninformative data and reducing analysis time. Data are compared at individual retention times and m/z values to reduce the rate of false positives, thus utilising as much information as possible from each LC-MS run. This method has been applied to high resolution and mass accuracy MS using FTICR-MS (Meng *et al.,* 2006) to measure the minimum fold change detectable and the accuracy of the relative ratio calculations. Much smaller changes in protein expression were detected of less than two fold, thus the method in combination with automated dMS software has potential application to complex mixtures for relative protein quantification. An alternative to using integrated extracted ion chromatograms for relative quantification without stable isotope labels is to use spectral count. This counts the number of spectral copies of peptides detected from a particular protein, or the number of tandem mass spectra produced from peptides derived from a particular protein, and relates it to the number of expected spectral copies from the observed proteins (Wong *et al.,* 2007). For this, it is assumed that the most abundant proteins are most likely to generate peptides in high abundance which will be selected for MSMS more frequently. Samples are analysed separately using the same data acquisition protocol, and the number of tandem mass spectra corresponding to each protein is normalised to account for protein length and the expected number of tryptic peptides (Nesvizhskii *et al.,* 2007). Using this method, protein abundance is linearly correlated with spectral count over a dynamic range of 100 (Liu *et al.,* 2004). In the same study, no correlation was found between percentage sequence coverage and number of peptides identified per protein for use as measure of relative abundance. These data suggest that for a global profiling experiment in which large differences in protein abundance are expected, spectral counting provides a reproducible and reliable strategy for relative quantification without stable isotope labels (Zhang *et al.,* 2006). This can be combined with

analysis of unknown proteins detected in MS where analysis is driven by differential protein expression of the same peptide from each sample using computational tools (Nesvizhskii *et al.*, 2007). This uses statistical information from differential MS signal intensity of the same peptide across multiple samples to target MS analysis for protein identification. This compensates for small sample sizes and the limited number of CID events that can take place at a given time, but some peptides may be common to more than one protein, thus introducing a degree of ambiguity. Label free LC-MSMS experiments are becoming more popular for comparative proteomics, with the advantage of relatively low cost (avoiding the use of stable isotopes) and the potential to compare multiple samples simultaneously. With significant advances in computational tools for statistical analysis and interpretation of data, this may become a highly efficient strategy for quantitative proteomics (Wong *et al.*, 2007).

***Label free relative quantification using mass tagging***

Relative quantification of proteins in two samples can be achieved by differential mass tagging. Other than methods of stable isotope incorporation, the derivatisation procedure of guanidination has been used prior to relative quantification (Cagney and Emili, 2002) prompting the term mass-coded abundance tagging (MCAT). Peptides from two different samples are digested with trypsin, prior to modification of one sample with *O*-methyisourea, guanidinating lysine residues. Samples are combined and analysed by reversed-phase LC-MS, upon which abundance of lysine peptides in both samples (detected with a mass shift of 42Da according to the modification of lysine to homoarginine) is calculated from LC trace intensities of extracted ion chromatograms. This approach targets a single amino acid residue for chemical modification to report on the abundance of peptides in different samples. However, modification by *O*-methylisourea alters the propensity of lysine peptides to ionise, thus the response of lysine and homoarginine terminated peptides in mass spectrometry cannot be used as a quantitative measure of relative abundance. This reaction is carried out on the pre-digested material, rather than on the intact protein, thus it is important to ensure complete digestion has taken place prior to the addition of *O*-methylisourea in order to use this method for quantification.

***Stable isotopes: differential labelling/derivatisation***

As discussed in section 1.2.4, stable isotope labelling is used in quantitative proteomics to assess relative protein abundance between two protein samples. This incorporates a mass shift between labelled and unlabelled analytes that can be detected by mass spectrometry. To

incorporate stable isotopes into relative quantification experiments, methods include metabolic labelling, incorporation during digestion, and through the use of isotopically modified chemical reagents and chemical tags (Figure 8). Differential labelling involves labelling one sample with a 'heavy', stable isotope labelled reagent, and another with a 'light' unlabelled equivalent. This type of labelling may be applied at the protein or peptide level, or during proteolysis of proteins to peptides, may be non-specific, or may target selected amino acid residues. As an example of non-specific differential labelling, proteins are digested with trypsin using either $H_2[^{18}O]$ or $H_2[O^{16}]$ as the buffer for digestion. This incorporates two [$^{18}$O] atoms into the C-terminus of all tryptic peptides allowing the distinction of labelled and unlabelled peptides from two populations (Yao *et al.,* 2001). This incorporation is stable under normal conditions for MS and tandem MS (Schnölzer *et al.,* 2005) and does not compromise peptide identification efficiency (López-Ferrer *et al.,* 2006). Alternative proteases differ in the way that [$^{18}$O] can be used; trypsin, chymotrypsin, endoproteinase Lys-C and endoproteinase Glu-C incorporate two [$^{18}$O] (+4amu) into peptides. Once proteolysis is complete, proteases continue to form esters reversibly with the C-terminal amino acid. Eventually all of the [$^{16}$O] is displaced, the rate of which depends on the C-terminal amino acid residue, peptide size, sequence and protease used. This method is not universal as a labelling strategy as the C-terminus cannot be labelled. Alternative *in vitro* labelling strategies incorporate stable isotope labels into analytical reagents, treating two samples differentially with 'light' and 'heavy' reagents. There are many examples of this technology including acetylation of two peptide mixtures, one with [$^1H_3$]acetyl groups and one with [$^2H_3$]acetyl groups (Ji *et al.,* 2000) in which primary amino groups at the N-terminus and the ε-amino of lysine are derivatised. When the samples are mixed and analysed by MS, the isotope ratio can be measured, allowing detection of the difference in relative concentration of that particular peptide. This is a global coding strategy that distinguishes lysine terminated tryptic peptides from those that are arginine terminated as they are labelled at both the N-terminus and the lysine residue, doubling the mass offset from the unlabelled peptide. Other primary amine derivatising agents, for example benzoate ester (Julka and Regnier, 2004) can also be labelled, for example with [$^{13}$C]. This strategy can also be adapted to label only the N-terminus of peptides, by blocking the ε-amino of lysine with succinic anhydride or guanidination; this has the added benefit of improving ionisation of these peptides. In addition, free amino groups are targets for differential labelling in the intact protein, using deuterated derivatising reagents prior to tryptic digestion and MS analysis (Schmidt *et al.,* 2005). Alkylation (intact proteins) and esterification (peptides) are also popular differential labelling targets, with the use of deuterated agents such as 4-vinylpyridine which alkylates all –SH residues following

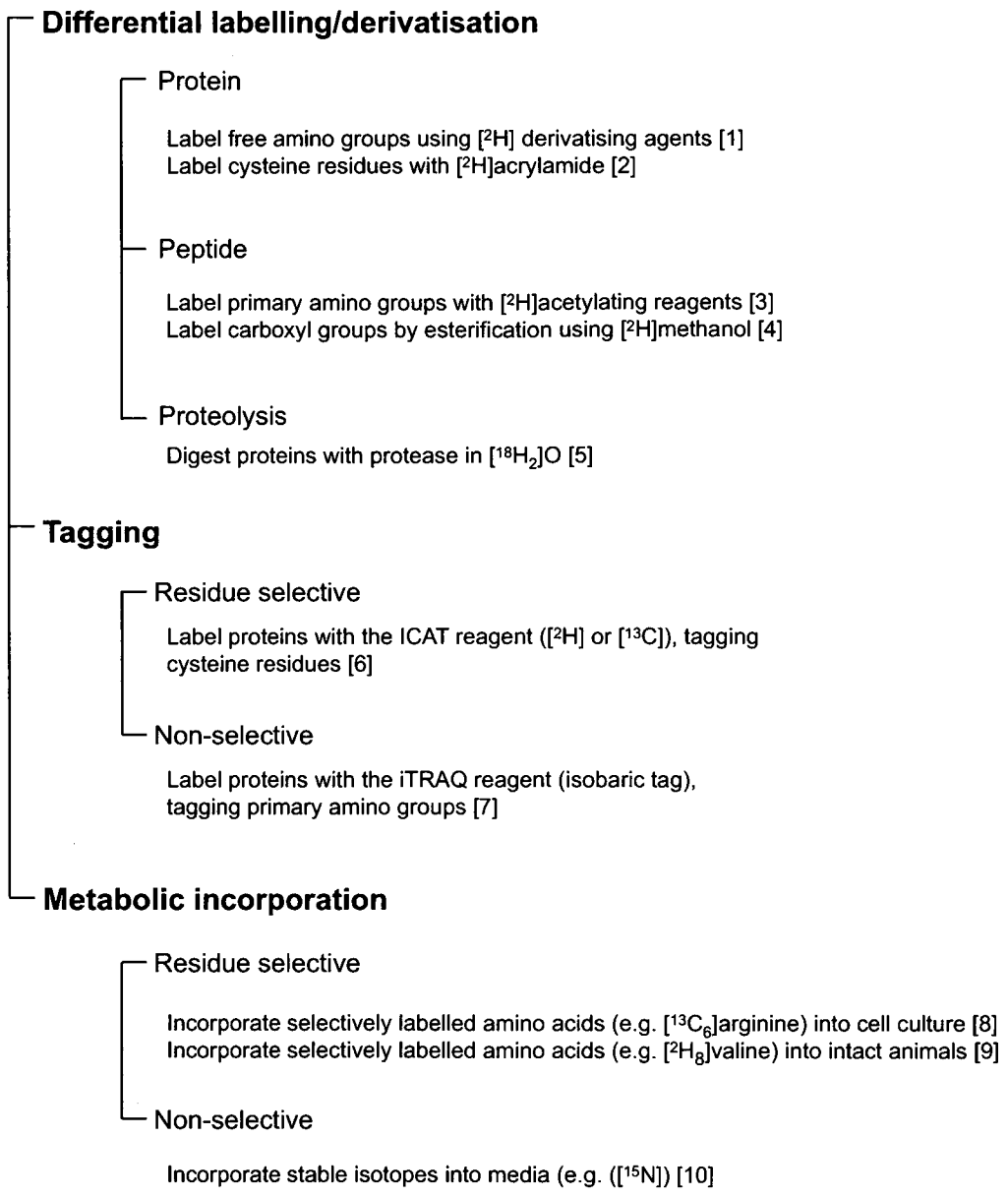



**Figure 8. Stable isotope labelling strategies for comparative proteomics.**
Examples of methods for incorporation of stable isotope labels into proteins and peptides for relative protein quantification by mass spectrometry are indicated. These fall into three main categories; differential labelling/derivatisation, tagging and metabolic incorporation. For discussion of the relative merits of these strategies, refer to text. References indicating examples of labelling strategies are given in each case.

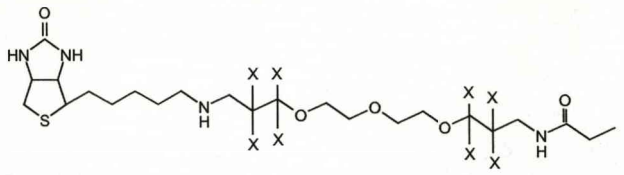

[1] Schmidt *et al.*, 2005
[2] Sechi and Chait, 1998
[3] Ji *et al.*, 2000
[4] Goodlett et al., 2001
[5] Yao *et al.*, 2001
[6] Gygi *et al.*, 1999
[7] Ross *et al.*, 2004
[8] Ong *et al.*, 2002
[9] Doherty *et al.*, 2005
[10] Snijders *et al.*, 2005

disruption of the disulphide bond (Sebastiano *et al.,* 2003) and methylation of C-terminal carboxyl groups of aspartic and glutamic acid using [$^{2}H_3$]methanol. Priorities for selecting the appropriate labelling strategy include the cost with which effective labelling can be achieved, the ease of label incorporation into target peptides, and detection in the analytical system of choice for relative quantification. To provide a global labelling strategy using analytical reagents including those discussed above, two or more labels may be incorporated. This combines the benefits of alternative methods, for example carboxyl groups labelled using [$^{18}O$] during proteolysis with primary amine groups labelled by acylation with deuterated acetoxysuccinamide (Liu and Regnier, 2002). Using this protocol, the number of primary amine and carboxyl groups will be determined from the mass shift relating to the heavy isotope; this allows the C-terminal peptide to be distinguished as it will only be labelled with [$^{2}H$] from acylation. However, several of these strategies are nonspecific and can result in incomplete labelling, not withstanding the need for high purity reagents. For complex mixtures of peptides that require chromatographic separation prior to mass spectrometric analysis, deuterated peptides can elute separately in reversed-phase, thus more expensive isotopes of carbon [$^{13}C$] and nitrogen [$^{15}N$] must be used, or data processing changed.

***Stable isotopes: tagging methods***

Based on a similar principle to MCAT, stable isotope labelled tags are synthesised that target specific amino acids in peptides and proteins. Amino acids such as cysteine are often used due to the specific chemistry of its sulphydryl group which can be exploited in both a differential labelling approach, for example using deuterated acrylamide to alkylate cysteine residues, or as the target for a specifically designed stable isotope labelled tag. This allows simplification of the peptide mixture and adds information about the peptide as to the number of the selected amino acid included, facilitating database searching and protein identification. Targetting and tagging specific amino acids for quantification is the basis of isotope-coded affinity tagging (ICAT; Gygi *et al.,* 1999) in which the samples are treated with isotopically light and heavy ICAT reagents. The ICAT reagent consists of a reactive group that is cysteine specific, a deuterium labelled linker region and a biotin tag for affinity chromatography and selective recovery of bound peptides (Figure 9). This is processed and peptides can then be separated and will be distinguished in MS due to the mass shift relating to the heavy label. After proteolytic digestion, the labelled peptides are affinity purified using a strepdavidin affinity matrix that binds the biotin tag, thereby achieving simplification of the peptide mixture at the same time as incorporating the isotopic label. The relative intensities of these peaks can then

a.




b.

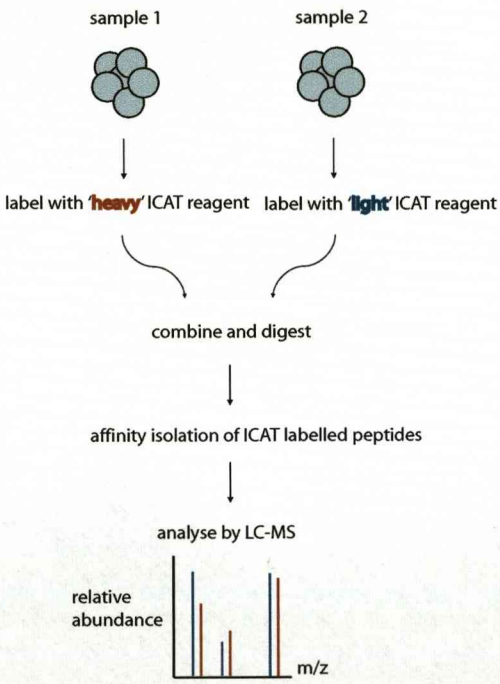



**Figure 9. Isotope coded affinity tagging (ICAT).**
Samples are treated separately with isotopically light and heavy ICAT reagents and digested with a protease, for example trypsin. The ICAT reagent (a) consists of a reactive group that is cysteine specific, a stable isotope labelled linker region and a biotin tag for affinity chromatography and selective recovery of bound peptides. For the 'heavy' reagent, X is replaced with the stable isotope, for example deuterium [$^2$H]. For relative quantification of protein abundance (b), the two samples once labelled are combined and digested, upon which peptides containing cysteine residues are isolated and analysed by LC-MS. Quantification is based on relative ion signal intensity of stable isotope label tagged peptides which can be distinguished by mass spectrometry according to the mass difference between 'heavy' and 'light' ICAT reagents.

be used for quantification of the amount of protein in the two cell preparations. Observed peak ratios for isotopic analogues are highly accurate as there are no chemical differences between the species and they are analysed in the same experiment. Although used successfully in combination with multidimensional chromatography to detect and quantify low abundance proteins (Gygi *et al.,* 2002), limitations of this method include non-specific binding to the strepdavidin affinity matrix, and multiple subsequent reactions at the same site. However, by the principle of internal standardisation, providing labelled and unlabelled samples are treated in an identical way, problems including selective capture of some cysteine containing peptides should be compensated for. This approach also limits the identification of post-translational modifications as well as the analysis of proteins and peptides containing no cysteine residues, as the affinity tag is cysteine specific. Quantification is also limited to MS only as the tag complicates fragmentation spectra, and deuterium isotopes do not permit the use of reversed-phase chromatography as the differentially labelled peptides would not co-elute. This has been improved through the use of cleavable ICAT reagents prior to MS analysis and co-eluting tags using [$^{13}$C] as the stable isotope label (Li *et al.,* 2003).

In an alternative strategy that is not selective for specific amino acid residues, isobaric tags for relative and absolute quantitation (iTRAQ) have been developed that target primary amino groups on the N-terminus and lysine side chains of peptides (Ross *et al.,* 2004). In principle, this is a strategy for relative quantification that may be applied to achieve absolute quantification if the isobaric tags are incorporated onto synthetic peptides (discussed in section 1.5.3). The tags are labelled and isobaric so that the peptide with incorporated tag is indistinguishable through chromatographic separation and MS but that generates a specific reporter ion in MSMS spectra. Four tags were initially developed, of mass 114, 115, 116 and 117Da, permitting the relative quantification of up to four proteins. Each tag contains a reporter group, a balance carbonyl group to compensate for the different mass of the reporters and a peptide reactive group (Figure 10). Once tagged, up to four samples are mixed and analysed by tandem MS upon which fragmentation of the tag attached to peptides gives rise to specific low molecular mass reporter ions, the intensity of which relates to the relative amount of the peptide in each sample. For this approach, peptide separation is important to avoid co-eluting species of similar mass that may compromise quantification. For this reason, the method has been combined with LC-MALDI offline for fragmentation by MALDI-ToF/ToF MSMS in addition to online separation prior to ESI-Q-ToF MS (Wiese *et al.,* 2007). In the same study, iTRAQ reagents were used to label intact proteins prior to initial separation by gel electrophoresis,

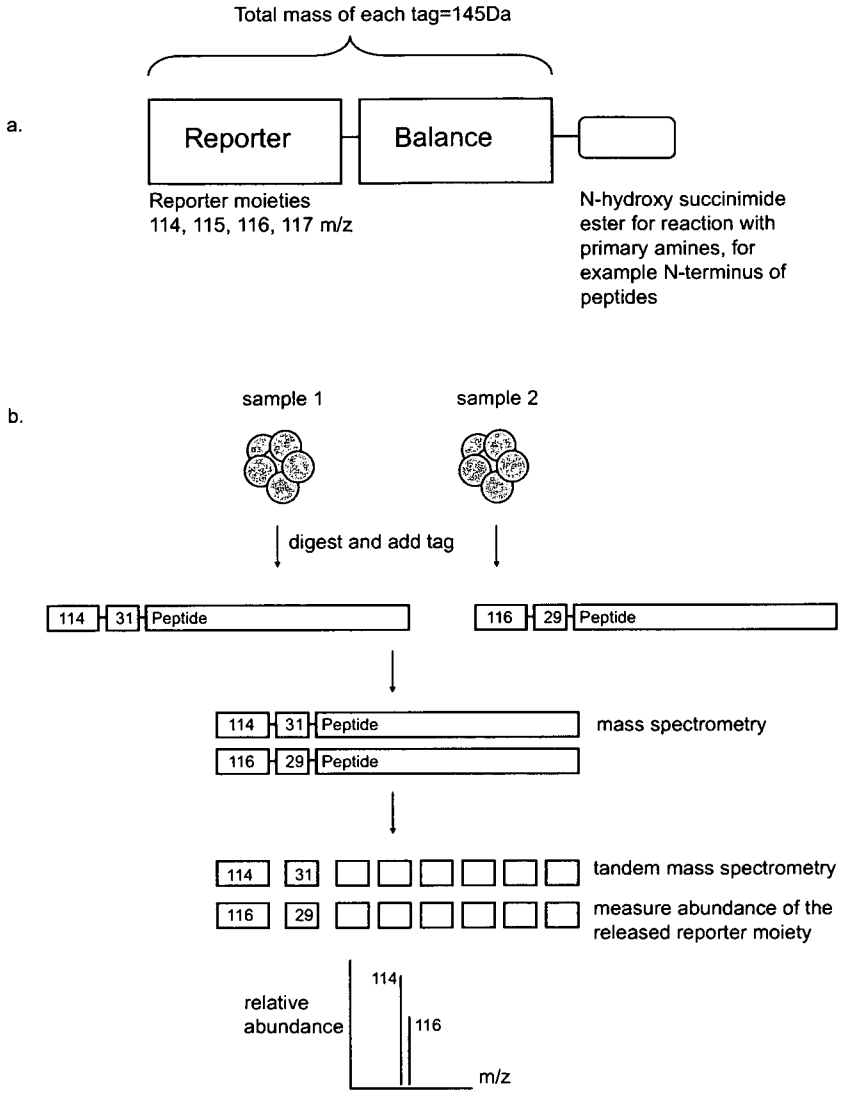



**Figure 10. Isobaric tagging for relative and absolute quantitation (iTRAQ).**
Samples are labelled with isobaric tags (a) such that the peptide with incorporated tag is indistinguishable through chromatographic separation and mass spectrometry but that generates a specific reported ion in tandem mass spectra. Tags contain a specific reporter moiety, a balance to maintain the same overall mass of the precursor ion and a reactive group to attach to primary amines, for example the N-terminus of peptides. For quantification (b), tagged samples are mixed and analysed by tandem mass spectrometry upon which fragmentation of the precursor ion gives rise to the specific low mass reporter ions, the relative signal intensity of which can be used to determine protein abundance.

proteolytic digestion and MSMS analysis. This approach is more time consuming as quantification is dependent on product ion scans, but this is matched with an increased peak capacity and signal to noise ratio due to quantification using fragment ions selected from specific precursors, providing non related peptide fragment ions are not present in the fragment ion spectrum. Consequently, the use of iTRAQ for relative quantification of proteins is highly sensitive with low technical variation when compared to biological variation (Gan *et al.*, 2007). This strategy is also reproducible across multiple sample injections for analysis by LC-MSMS used to maximise number of protein identifications, when combined with quantification in a single experiment (Chong *et al.*, 2006). In a similar approach avoiding the use of isobaric tags, amino acids may be modified using stable isotopes in such a way that the peptide will have the same overall mass but a different incorporated isotope, for example one is labelled with [$^{13}C_1$] and the other [$^{15}N_1$]. This will be distinguished in the immonium ion for the selectively labelled amino acids upon tandem MS. The disadvantage of this approach is that a mass difference of 1Da may be difficult to distinguish (Julka and Regnier, 2004).

***Stable isotopes: metabolic incorporation***

Internal standards can be incorporated into the system of interest *in-vivo* by metabolic incorporation of the stable isotope. This can be achieved using stable isotope labelled precursors, for example [$^{15}$N]$H_4Cl$ as the sole nitrogen source in the media, incorporating [$^{15}$N] into every peptide during synthesis. Proteins differentially labelled in this way are subsequently mixed, subjected to proteolytic digestion (with or without prior protein separation, for example by gel electrophoresis) and analysed by mass spectrometry (for example, Snijders *et al.*, 2005). Alternatively, labelled, essential amino acids can be added to amino acid deficient cell culture media (Stable Isotope Labelling by Amino Acids in Cell Culture, SILAC; Ong *et al.*, 2002), and are consequently incorporated into all proteins as they are synthesised (Figure 11). In this way, the entire population of proteins can be fully labelled at the start of the experiment for quantitative comparison with unlabelled populations upon mixing, proteolytic digestion and MS. This eliminates the need for chemical labelling and affinity purification, and the method is compatible with virtually all cell culture conditions. Labelling with [$^{15}$N] has limitations, in that labelling is not uniform; the number of [$^{15}$N] atoms introduced into each tryptic peptide will vary depending on the amino acid content. It is also more difficult and expensive to make [$^{15}$N]-substituted media for mammalian cell culture (Ong *et al.*, 2002). It is consequently beneficial to use stable isotope labelled amino acids such as [$^{13}C_6$]arginine and [$^{13}C_6$]lysine so that each tryptic peptide contains a single stable isotope, resulting in a constant mass offset between

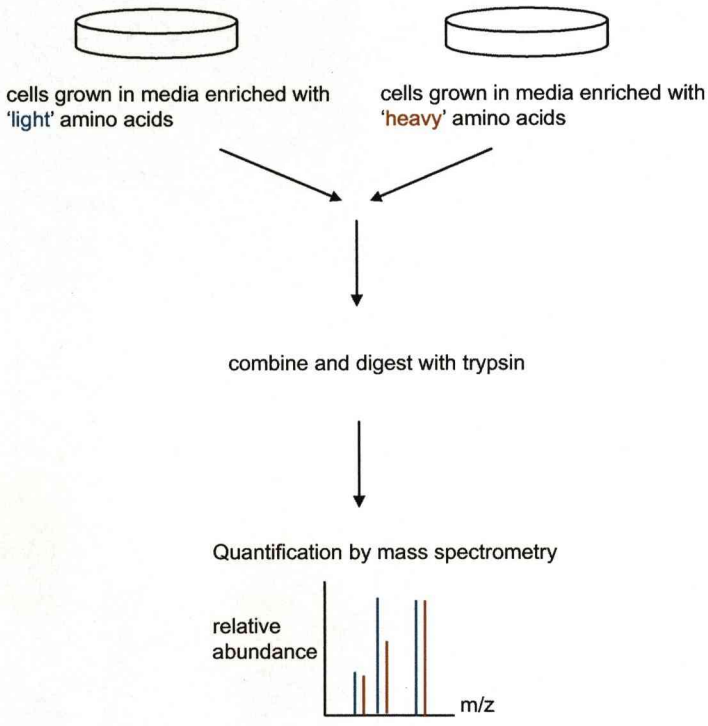



**Figure 11. Stable isotope labelling by amino acids in cell culture (SILAC).**
Internal standards can be incorporated into the system of interest *in-vivo* by metabolic incorporation of a stable isotope. This can be achieved using stable isotope labelled precursors, for example $[^{15}N]H_4Cl$ as the sole nitrogen source in the media, or labelled, essential amino acids can be added to amino acid deficient cell culture media, and are consequently incorporated into all proteins as they are synthesised. Material labelled in this way can be combined with that grown in the presence of the 'light' isotope, digested with trypsin and analysed by mass spectrometry. The mass shift relating to the 'heavy' isotope label can be distinguished by mass spectrometry, and the relative signal intensity of 'heavy' and 'light' isotopes can be used as a measure of protein abundance.

analyte and standard of 6Da. It is important when designing a SILAC experiment that the amino acid is abundant, essential for growth and is not likely to be altered by further processing, for example deamidation of asparagine to aspartic acid. For animal studies, the amino acid used must be essential (cannot be synthesised by the organism) and of high natural abundance to ensure labelling of the majority of peptides for quantification (Beynon and Pratt, 2005). Abundant amino acids such as leucine are often used, labelling approximately 70% of tryptic peptides (Ong *et al.,* 2002) and readily available in deuterated form. Leucine labelling *in vivo* is also a useful aid to peptide mass fingerprinting for protein identification as the mass shift between unlabelled and labelled versions of the peptide relates to the number of leucine residues in the peptide (Pratt *et al.,* 2002). However, deuterium labelling complicates reversed-phase separations, so this is not always the most appropriate strategy. Stable isotope labels can also be introduced at the whole animal level to determine rates of protein turnover, for example by feeding [$^{15}N$] enriched diets (Wu *et al.,* 2004), or deuterated amino acids, for example [$^{2}H_8$]-valine (Doherty *et al.,* 2005). The difficulty with this approach is that the precursor is subject to metabolism which can cause redistribution of label as degradation of pre-existing proteins will also contribute to the pool of unlabelled amino acids. The extent of this pool dilution effect can be calculated and is stable over an extended labelling period, thus allowing it to be taken into account when measuring rates of protein synthesis and degradation from incorporation of stable isotope label into the protein pool (Doherty *et al.,* 2005).

As for all quantitative proteomics experiments, regardless of strategy, it is essential that data produced are reliable, accurate and reproducible across groups, laboratories and time. To make significant advances in systems biology, particularly for biomarker discovery, quantitative data must be comparable across multiple platforms and in particular between different research institutions and groups with access to alternative instrumentation and resources across the world. To this end, 'The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study' was designed in which 52 participants generated relative quantification data (with and without the employment of stable isotopes) for eight proteins in two mixtures using a variety of techniques (Turck *et al.,* 2006). Although insufficient data sets were obtained to draw statistically significant conclusions, this study highlighted the range of results that can be achieved using alternative relative quantification strategies. In particular, variation between data sets, although more consistent for MS based approaches than gel electrophoresis based approaches was considerable. This was also a useful exercise to report the need for additional, more reliable methods to report and compare quantitative proteomics data. This was

particularly true for label free quantitative methods, where internal standards are not incorporated to measure variation between sample runs, differences in sample collection, preparation protocols, experimental design, platform stability and sample stability have significant impact on the reproducibility of results, especially as these are most likely to vary between laboratories. To compensate for these, algorithms have been designed to compare the similarity of mass spectral data, based on resolution, signal intensity, elution profile and signal to noise ratio, for example 'Chaorder' (Prakash *et al.,* 2006, Prakash *et al.,* 2007). This calculates an alignment score with potential application to alternative LC-MS platforms, including choice of instrument and chromatography column to assess reproducibility, thus reducing bias in comparative proteomics experiments. This technology provides a significant advantage to LC-MS analyses for quantitative proteomics in which the use of stable isotopes is undesirable, or impractical, but standardisation between experimental platforms is essential.

### 1.5.3 Absolute quantification

Absolute quantification in proteomics defines the number of molecules of a particular protein, rather than relative quantification which defines protein abundance in relation to another protein, or the same protein in another sample. In principle, any of the approaches adopted for relative quantification may also be used for absolute quantification if appropriate reference standards are available for all analytes in known amounts.

#### *Absolute quantification using western blotting*

Measurement of protein abundance using non stable isotope labelling methods has previously been discussed in the context of relative protein quantification (section 1.5.2). To move from relative to absolute quantification avoiding the use of stable isotope labelling, technologies for western blot analysis of protein abundance (described in section 1.5.2) use the more sensitive technique of fluorescence. Proteins separated by gel electrophoresis and transferred to a nitrocellulose membrane are probed with primary antibody, followed by secondary antibody carrying a fluorescent label. These labels are excited by light and detected, for example using flow cytometry to detect proteins fused with green-fluorescent protein (GFP; Bar-Even *et al.,* 2006). To measure absolute protein levels, epitope tagged fusion proteins are added to an organism's genome, and proteins containing these fusion tags are subsequently expressed under normal conditions, controlled by their natural promoters (Ghaemmaghami *et al.,* 2003). Proteins containing the fusion tags are detected using fluorescence microscopy and quantified using known amounts of antibodies specific to the epitope contained on the tag. These data

can be compared to mRNA abundance and translation rate for comprehensive analysis of protein metabolism (Belle *et al.*, 2006).

With increasing knowledge of the peptides most likely to be observed upon MS based on ionisation efficiency, retention time during a reversed phase separation and other chemical and physical properties (Mallick *et al.*, 2007), some of the label free strategies previously discussed for quantification may also predict absolute protein abundance. By comparing the number of peptides detected from a particular protein with its abundance, and correcting for the probability of observing each of these based on their composition, absolute protein expression measurement may be possible (Lu *et al.*, 2007). Protein abundance measured in this way agrees with alternative techniques avoiding stable isotope labels, including western blotting, flow cytometry and 2D SDS-PAGE. Data also correlate well with mRNA abundances, allowing measurements of mRNA directed gene expression regulation.

***Stable isotope labelling for absolute quantification***

To use stable isotopes for absolute quantification of individual proteins, the true internal standard would be the corresponding protein, expressed in pure and stable isotope labelled form and quantified (Brun *et al.*, 2007). This is challenging on many fronts, including the expression of a native protein in a heterologous system to effect labelling, purification of the protein, and subsequent mass spectrometric analysis of the complex isotopic profile of the analyte and standard protein. Rather than adopt a protein-based approach, a strategy for absolute quantification using proteotypic peptides as surrogates for the protein of interest has emerged, employing stable isotope labelled peptide internal standards as 'signature' or 'proteotypic' peptides, chemically synthesised and incorporating stable isotopes (Figure 12). This approach initially used a known amount of a chemically synthesised, deuterated peptide from apolipoprotein A1 to quantify the absolute amount of the purified protein (Barr *et al.*, 1996) and was superseded using [$^{13}$C] or [$^{15}$N] labelling of specific amino acids for absolute quantification of proteins in complex mixtures, in addition to those that have been phosphorylated (Gerber *et al.*, 2003) and later ubiquitinated (Kirkpatrick *et al.*, 2005) post-translationally. Internal standard peptides are added to the protein mixture in a known amount during proteolytic digestion allowing the ratio of signal intensity between standard and analyte in MS to dictate the abundance of the analyte. This method has been widely applied, for example to clinical samples in which protein expression changes with disease can be used to predict diagnostic biomarkers (Kuhn *et al.*, 2004) and to cells in culture allowing absolute

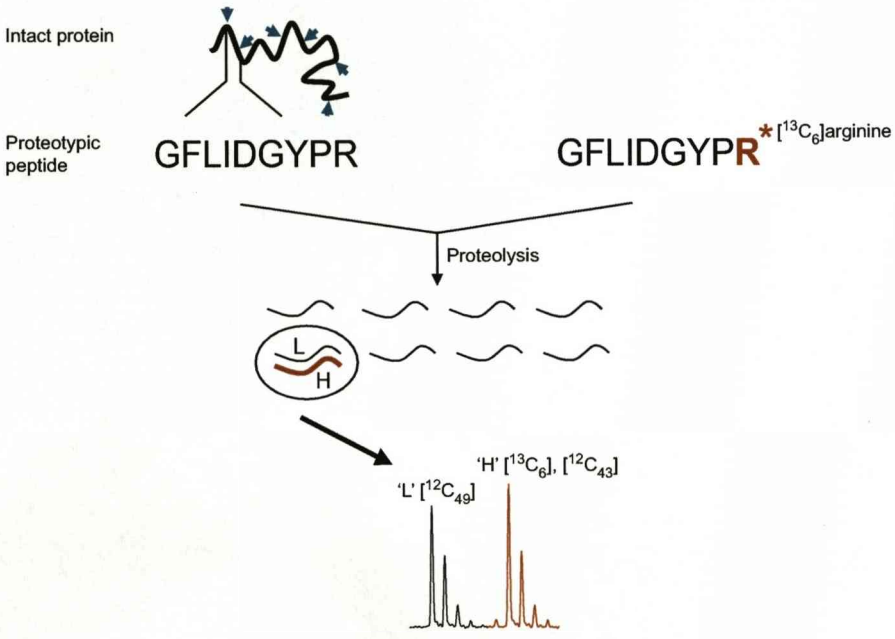



**Figure 12. Internal standardisation for absolute protein quantification.**
For absolute quantification of proteins, a surrogate peptide is selected and an identical copy is synthesised with an incorporated stable isotope label. The synthetic peptide internal standard is added to digested protein material in a known amount. Upon mass spectrometry analysis, the stable isotope labelled internal standard peptide will be distinguished from the surrogate peptide from the analyte by the mass shift relating to the 'heavy' label. The relative signal intensity of the two peptide ions can be used as a measure of absolute abundance of the analyte protein as the two peptides will respond in exactly the same way within the mass spectrometer.

quantification of proteins in animal tissues (Ishihama *et al.,* 2005). Synthetic peptide approaches have also been used to adapt previous quantitative methods, for example iTRAQ can be used for absolute quantification if synthetic, stable isotope labelled peptides are tagged with the iTRAQ reagents (Ross *et al.,* 2004). These studies highlight the importance of versatility of this methodology to alternative biological samples and mass spectral detection systems. However, to quantify multiple proteins, each requires at least one stable isotope labelled peptide that must be independently synthesised at relatively high cost. Moreover, each peptide must be separately purified and quantified (Pan *et al.,* 2005).

To streamline this approach, a novel strategy has been introduced as an efficient alternative to the chemical synthesis of multiple stable isotope labelled peptides (Beynon *et al.,* 2005, Pratt *et al.,* 2006). The 'QconCAT' method uses artificial genes, designed *de novo* to direct the synthesis of novel proteins which are assemblies of signature standard peptides (Q-peptides), derived from a number of discrete proteins. Usually, these Q-peptides are arginine or lysine terminated at the C-terminus, as they represent tryptic peptides derived from digestion of the analyte proteins. The artificial **q**uantification **concat**amer (QconCAT) protein contains the Q-peptides appropriately flanked with added features including an initiator codon, a purification tag and protective sacrificial regions. The gene is transformed into and expressed in a heterologous system, usually bacterial. The expression strain is grown in chemically defined media, uniformly isotopically labelled (for example, using $[^{15}N]H_4Cl$ as the sole nitrogen source) or containing specific stable isotope labelled amino acids at a high isotope enrichment, such that the artificial protein becomes fully labelled (Figure 13). The artificial QconCAT protein is purified by virtue of the affinity tag and quantified using a suitable procedure (Pratt *et al.,* 2006). The QconCAT protein is added to a complex mixture of analyte proteins, and subsequent proteolysis releases both the stable isotope labelled standard and the cognate peptide from the analyte. The known quantity of standard added can then be used for absolute quantification of the analyte. Since quantification of the QconCAT protein will define in absolute terms the quantity of each of the surrogate peptides, the QconCAT strategy provides an efficient means to multiplex absolute quantification. Tryptic peptides are typically 10-15 amino acids long, thus proteotypic Q-peptides from 50 proteins could be encoded in a protein comprising 500-750 amino acids. To quantify 50 proteins at one Q-peptide per protein, a QconCAT strategy is about 15% of the cost of comparable synthetic peptides, and would yield about 250nmol of protein compared to 5nmol of each synthetic peptide. The Q-peptides are present, by design, in

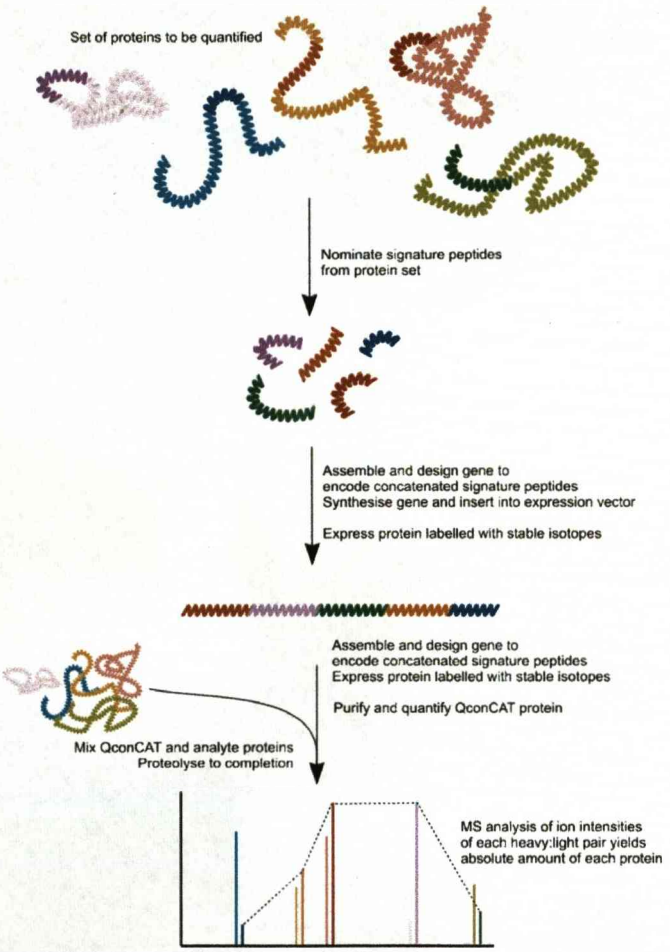



**Figure 13. The QconCAT strategy for absolute quantification.**

Taken from Pratt *et al.*, 2006

stoichiometrically known amounts (usually equimolar), so that each analyte peptide (and therefore protein) is simultaneously quantified.

### 1.5.4 Challenges for absolute quantification using surrogate peptides

#### *Proteotypic peptides*

Using a synthetic peptide approach requires a great deal of consideration into peptide selection and validation, particularly for the analysis of complex protein mixtures in terms of retention time by reversed-phase chromatography, ionisation efficiency and fragmentation if using a tandem MS approach. Proteotypic peptides are selected based on previous detection within the analytical system of choice. These are typically peptides that ionise well in MS (Baumgart *et al.,* 2004), are unique and relatively abundant (within the analyte system of choice). It is also important that these peptides behave in the same way in a complex of mixture of proteins in which they will need to be detected for absolute quantification (Pan *et al.,* 2005). Peptides with certain physicochemical properties have an increased chance of being detected by MS, and this information can be used to develop computational tools to predict proteotypic peptides (Mallick *et al.,* 2007), of significant benefit to design of a QconCAT experiment. The order of amino acids in each peptide may also influence the probability of preferential detection, thus creating many combinations of varying significance. With increasing sharing of data in public repositories from more directed analyses, prediction of proteotypic peptides for proteomics will be a valuable tool for absolute quantification with and without stable isotope labelling.

#### *Complete proteolysis*

Q-peptides are concatenated in the QconCAT protein disconnected from their normal primary sequence context, for example T3 from adenylate kinase exists in the QconCAT protein as VIR↓**GFLIDGYPR**↓VVL, and in the native protein as TSK↓**GFLIDGYPR**↓EVK. This different context could influence quantification, as the abundance of a peptide released by proteolysis is used to report on the native protein (Kito *et al.,* 2007). However, this can only occur if either the QconCAT or the analyte proteins are incompletely digested, such that the yield of each peptide is incomplete and the main determinant of the rate of proteolysis of native proteins is primary sequence context, not higher order structure. Tightly folded proteins, particularly those with a high proportion of beta sheet, are intrinsically resistant to proteolysis (Hubbard *et al.,* 1998; Wu *et al.,* 1999). However, it is not expected that QconCATs would adopt tightly folded structures as they have no biological function (Pratt *et al.,* 2006). By contrast, unless care is taken in the prior denaturation of analyte proteins, their higher order structure would almost certainly

influence proteolysis, and could impact on absolute quantification. Incomplete analyte digestion is as much an issue for quantification using synthetic peptides as those using QconCATs.

## 2. Aims and Objectives

## 3. The QconCAT strategy for absolute quantification of chicken skeletal muscle soluble proteins

3.1 Changing proteome dynamics in chicken skeletal muscle ..... 44

3.2 QconCAT design ..... 45

3.3 Deployment of the QconCAT strategy ..... 46
3.3.1 Reliability of a QconCAT method ..... 46
3.3.2 Proteolysis ..... 47

3.4 Additional applications of QconCAT ..... 47
3.4.1 Using QconCAT to quantify skeletal muscle proteins from other species ..... 47
3.4.2 Quantification of normalisation using Equalizer™ beads ..... 47
3.4.3 Absolute quantification of the post-translation modification, deamidation ..... 48

## 2. Aims and Objectives

The aim of this research was the deployment of a QconCAT method for the absolute quantification of soluble proteins in chicken skeletal muscle during growth from 1d to 30d post hatch in both broiler (meat producing) and layer (egg producing) strains. This required purification of expressed protein (expression and labelling of QconCAT protein was courtesy of Dr. D.M. Simpson), co-digestion with analyte proteins, simultaneous detection of analyte and internal standard peptides by MS, and quantification. Through the execution of this study, the objective was to provide a rigorous test of the QconCAT method for absolute quantification of multiple proteins, incorporating an in-depth assessment of variance within this system.

## 3. The QconCAT Strategy for Absolute Quantification of Chicken Skeletal Muscle Soluble Proteins

### 3.1 Changing Proteome Dynamics in Chicken Skeletal Muscle

As the demand for quality poultry produce has risen, the drive to produce fast growing, lean birds has increased. This is reflected in the dramatic difference in growth rate and consequent size between birds selected for meat production (broilers) and those selected for egg production (layers; Figure 14). This difference is particularly well documented in the pectoralis (breast) muscle which increases in overall size as a percentage of total body weight at a far greater rate in broilers, as does the concentration of protein in the tissue. In fact, the increased selection pressure on these meat producing birds has led to disproportionate growth of the pectoralis muscle compared to overall body weight, comprising approximately 10% of total body weight at 45 days of age. By comparison, the pectoralis muscle of layer birds comprises only 5-6% of total body weight throughout development (Flannery *et al.*, 1992). Such a high growth rate places abnormal and extreme strain on the natural growth and development of the bird, which in many cases cannot be sustained. This causes bone abnormalities with repercussions for general health including pulmonary hypertension and ascites (Griffin and Goddard, 1994). The amount of poultry meat available to consumers in the United Kingdom is continuing to rise, from 1.5 million tonnes in 1995 to 1.7 million tonnes in 2002 (www.publications.parliament.co.uk). As a result, meat producing birds are bred to gain muscle at a very high rate compared to layers. This has serious welfare implications with extra weight placed on bones and joints causing lameness and joint infection. In addition, broilers that are unable to walk are more susceptible to pressure sores and skin infections from lying in

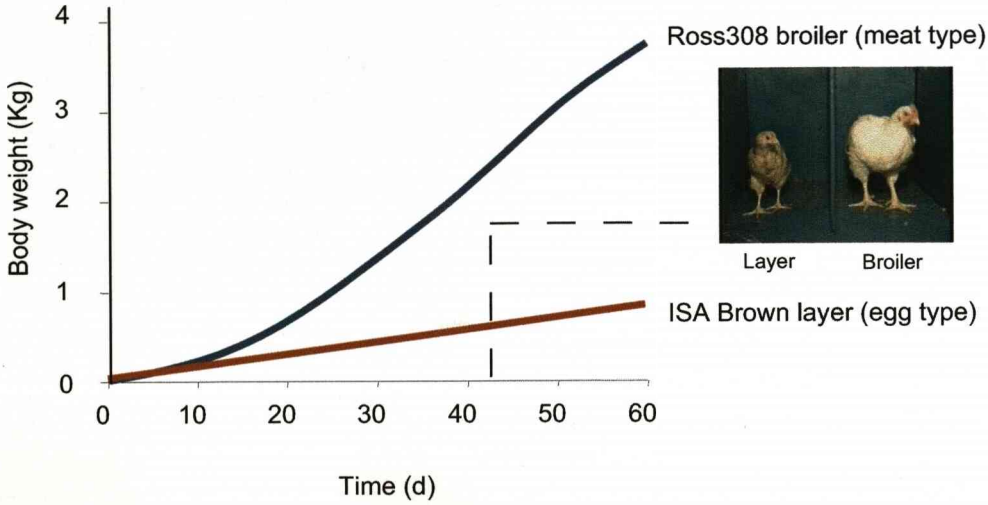



**Figure 14. Growth rates of broiler and layer chickens.**
Differential growth rates of broiler (meat producing) and layer (egg producing) chickens measured as overall body weight (kg) from 0 to 60 days. The inserted photograph depcits the dramatic difference in size of a broiler and layer bird at six weeks of age.
Data supplied by Heather McCormack (Roslin Institute).

excrement. Heart disease and respiratory problems are also common as the additional requirement for growth and metabolism increases the demand for oxygen, placing abnormal strain on the heart and respiratory system (www.chickens.rspca.org.uk).

To take control of the physiological effects of increased growth of the pectoralis muscle in broiler chickens compared to layers, an understanding of the composition and dynamics of the proteome of the pectoralis muscle is required. The pectoralis muscle of an adult chicken comprises predominantly white, type II glycolytic fibres, generating ATP by glycolysis. The soluble fraction of a pectoralis muscle homogenate predominantly contains glycolytic enzymes which are sufficiently abundant to dominate most proteomic analyses (Doherty *et al.,* 2004). Subsequently, these may be the most important proteins driving metabolic processes, including growth. For this reason, the pectoralis muscle provides a relatively simple tissue with which the soluble fraction can be used to identify specific changes in muscle composition responsible for the dramatic difference in growth rates of the two strains of chicken.

### 3.2 QconCAT Design

The QconCAT used in this study was designed as a quantification standard for abundant soluble proteins in chicken skeletal muscle. This incorporated the most abundant proteins from a muscle homogenate when run on a 1D gel, with one Q-peptide per analyte protein (Table 3, Figure 15). Each protein to be quantified was digested in-gel with trypsin and analysed by MALDI-ToF MS to identify peptide candidates for incorporation into the QconCAT protein. Selection of peptides, where possible, was based on propensity to ionise (those that gave the strongest signal on MALDI-ToF MS analysis) and mass range (between 1000 and 2000Da) ensuring that each was unique within the set for absolute quantification. The DNA sequence of each chosen peptide was concatenated into the QconCAT gene, with short, N- and C-terminal extension sequences to protect the true Q peptides for quantification. These included an initiator methionine residue and a C-terminal HisTag for purification of the protein (Figure 15). The QconCAT gene was subsequently expressed in *E.coli*, unlabelled, labelled with [$^{15}$N] using [$^{15}$N]$H_4Cl$ enriched media, or with incorporated, essential, labelled amino acids [$^{13}C_6$]arg/[$^{13}C_6$]lys (three separate expression strains), from which proteins were expressed in inclusion bodies (Pratt *et al.,* 2006). Proteins were separately purified and quantified for use as a set of internal standard peptides upon co-digestion with chicken skeletal muscle soluble proteins. Within the QconCAT protein, each peptide is in a strict stoichiometry so that the entire

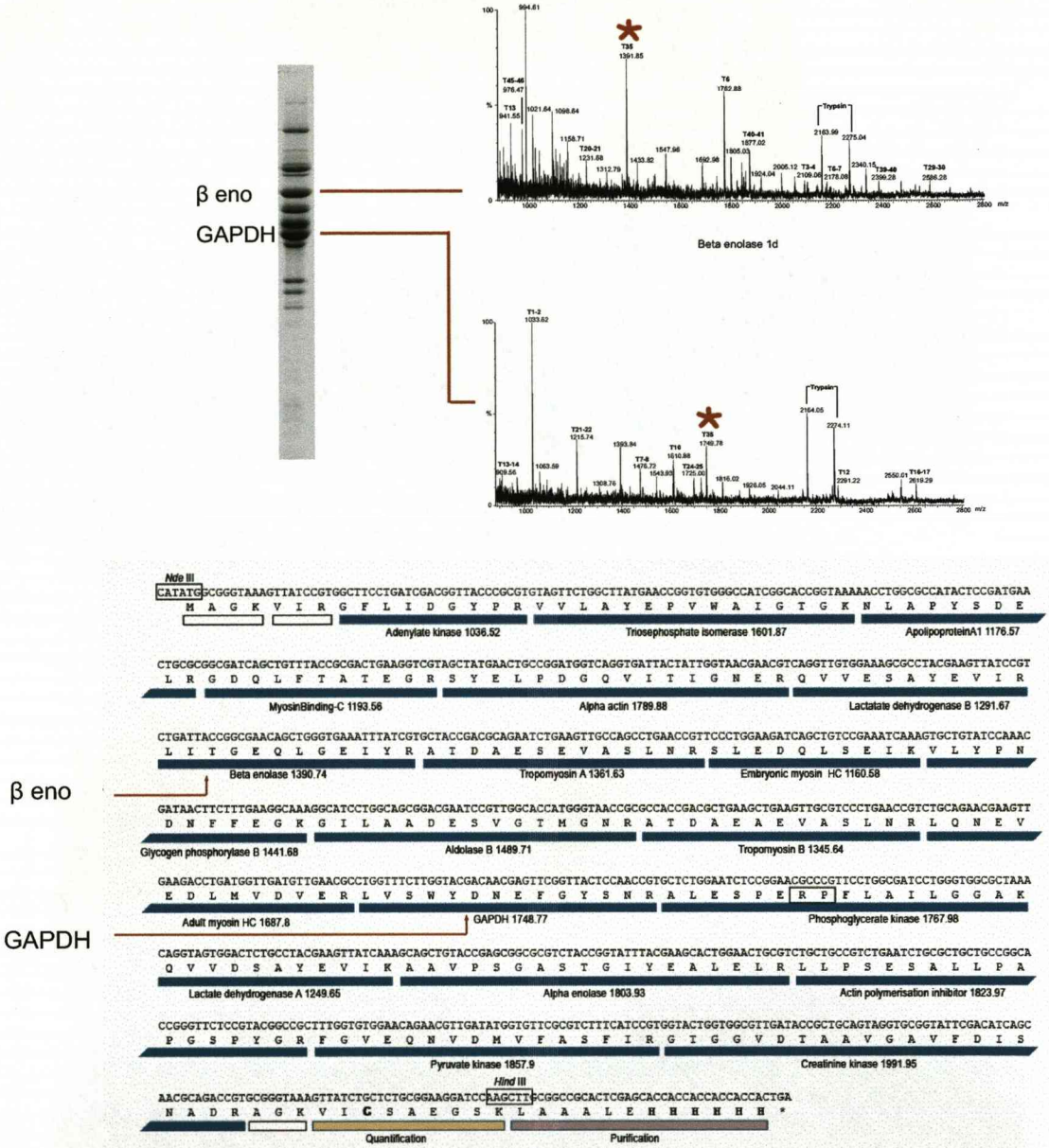



**Figure 15. Selection of QconCAT peptides and subsequent design of QconCAT gene constructed for the absolute quantification of chicken skeletal muscle soluble proteins.**
The most abundant proteins analysed by 1D SDS-PAGE were digested in-gel with trypsin prior to analysis by MALDI-ToF MS. Abundant, well ionised, unique peptides within the mass range 1000-2000Da were selected for incorporation into the QconCAT protein. These were assembled into the artificial QconCAT gene which was expressed in *E.coli* and labelled with stable isotopes for use as a set of internal standards for absolute quantification upon co-digestion with analyte proteins. Two examples of proteins selected for incorporation into the QconCAT are beta enolase (β eno) from which the tryptic peptide at 1390.85Da was used as a surrogate, and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) from which the tryptic peptide at 1748.77Da was selected. The DNA sequence, translated protein sequence (amino acids) and peptides to be generated by proteolysis with trypsin (blue bars) are indicated. The mass of each peptide, and the protein for which they are used as surrogate internal standards is given below with peptides from β eno and GAPDH indicated by a red arrow. particular features of the QconCAT protein include a non-cleavable -ArgPro- site (boxed), N- and C- terminal extensions to protect the true Q-peptides (white bars), a purification His Tag (purple bar), and a single cysteine residue for quantification (yellow bar). Adapted from Beynon *et al.,* 2005.

| Peptide | Sequence | Parent protein | Mass (Da) |
|---|---|---|---|
| T1 | MAGK | Construct, sacrificial | 405.2 |
| T2 | VIR | Construct, sacrificial | 386.25 |
| T3 | GFLIDGYPR | Adenylate kinase (AK) | 1036.52 |
| T4 | VVLAYEPVWAIGTGK | Triose phosphate isomerase (TPI) | 1601.87 |
| T5 | NLAPYSDELR | Apolipoprotein A1 (ApoA1) | 1176.57 |
| T6 | GDQLFTATEGR | Myosin binding protein C (MBC) | 1193.56 |
| T7 | SYELPDGQVITIGNER | Beta actin (β actin) | 1789.88 |
| T8 | QVVESAYEVIR | Lactate dehydrogenase B (LDH B) | 1291.67 |
| T9 | LITGEQLGEIYR | Beta enolase (β eno) | 1390.74 |
| T10 | ATDAESEVASLNR | Tropomyosin A (TM A) | 1361.63 |
| T11 | SLEDQLSEIK | Embryonic myosin (E myo) | 1160.58 |
| T12 | VLYPNDNFFEGK | Glycogen phosphorylase (GP) | 1441.68 |
| T13 | GILAADESVGTMGNR | Aldolase B (Aldo B) | 1489.71 |
| T14 | ATDAEAEVASLNR | Tropomyosin B (TM B) | 1345.64 |
| T15 | LQNEVEDLMVDVER | Adult myosin (A myo) | 1687.8 |
| T16 | LVSWYDNEFGYSNR | Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) | 1748.77 |
| T17 | ALESPERPFLAILGGAK | Phosphoglycerate kinase (PGK) | 1767.98 |
| T18 | QVVDSAYEVIK | Lactate dehydrogenase A (LDH A) | 1249.65 |
| T19 | AAVPSGASTGIYEALELR | Alpha enolase (α eno) | 1803.93 |
| T20 | LLPSESALLPAPGSPYGR | Actin polymerization inhibitor (API) | 1823.97 |
| T21 | FGVEQNVDMVFASFIR | Pyruvate kinase (PK) | 1857.9 |
| T22 | GTGGVDTAAVGAVFDISNADR | Creatine kinase (CK) | 1991.95 |
| T23 | AGK | Construct | 274.15 |
| T24 | VICSAEGSK | Construct, quantification | 892.42 |
| T25 | LAAALEHHHHHH | Construct, purification tag | 1408.68 |

Table 3. Peptides selected from chicken skeletal muscle soluble proteins represented in the QconCAT protein.

set of Q-peptides can be quantified in molar terms by determination of the QconCAT protein concentration.

## 3.3 Deployment of the QconCAT Strategy

For internal standardisation, the QconCAT protein was added to chicken skeletal muscle soluble fraction in a known amount prior to proteolytic digestion with trypsin. Upon analysis by MALDI-ToF MS, the mass spectrometric intensities of the 'heavy' stable isotope labelled peptide and the unlabelled skeletal muscle peptide in the tissue sample were used to accurately quantify each protein. This was conducted to assess the QconCAT method through the quantification of changes in protein expression during development of chicken skeletal muscle from 1d to 30d of growth in both broiler and layer strains. Through the full deployment of this strategy, sources of variation were established and assessed.

### 3.3.1 Reliability of a QconCAT method

For a reliable QconCAT method to achieve absolute quantification of proteins, it is imperative that the sources of variation and error are assessed and controlled. For quantitative proteomics experiments in which proteins are classified as differentially expressed across multiple samples, relative variance boundaries must be assessed and applied to distinguish significant changes in protein abundance (for example, Pandhal *et al.*, 2007). For the QconCAT method in particular, the difference between variance due to analytical replication and that between individual animals has defined the most likely source of variation. Alternative forms of ionisation in combination with reversed phase chromatography have also been tested as a measure of reproducibility, assessing robustness of this method across multiple platforms. As this is the first method to achieve absolute quantification of multiple proteins in a single experiment, it is important to test the accuracy of the method, particularly as compared to other methods for quantification. For a single protein, the QconCAT method has been compared to absolute quantification using a synthetic peptide incorporating a stable isotope label across the 30d of growth for the broiler strain. Quantification has also been compared to gel based approaches using densitometry of stained protein bands analysed by 1D SDS-PAGE in addition to intact mass analysis of proteins in the same sample as measured using ESI. However, as a true assessment of accuracy, a purified protein that is represented in the QconCAT for chicken skeletal muscle was spiked into a muscle sample in a known amount and subsequently quantified using QconCAT.

### 3.3.2 Proteolysis

Completeness of digestion has been quantified for both unlabelled standard and analyte proteins using completely pre-digested, labelled QconCAT peptides. This analysis has monitored the disappearance of intact proteins by 1D SDS-PAGE and appearance of limit peptides by MALDI-ToF MS. For analyte proteins where higher order structure is thought to be the major cause of incomplete proteolysis, denaturation of the protein by heating has also been investigated, as has the addition of a small amount of organic solvent to improve the activity of trypsin.

## 3.4 Additional Applications of QconCAT

### 3.4.1 Using QconCAT to quantify skeletal muscle proteins from other species

Relative quantification using differential stable isotope labelling of two species has been achieved using shared peptides, analysed by MS following proteolytic digestion (Snijders *et al.*, 2007). In the same study, *in silico* analysis of occurrence of shared peptides in multiple species revealed that, for example 30-50% of peptides are shared between *Mus musculus* and *Homo sapiens*, offering great potential for cross species quantification using peptide-based strategies discussed in section 1.5. Soluble muscle proteins are highly conserved amongst several species, sharing many peptides (although not necessarily those incorporated into the QconCAT) and providing an ideal opportunity to test the QconCAT method for quantification of peptides with subtle amino acid differences. This may have potential benefit to other commercially important species, allowing quantification of several proteins under different physiological conditions, for example in diseased states. To investigate this, readily available samples of mouse and carp soluble muscle proteins were analysed by 1D SDS-PAGE and quantification was achieved using MALDI-ToF MS and LC-ESI Q-ToF MS with chicken muscle QconCAT as the internal standard.

### 3.4.2 Quantification of normalisation using Equalizer™ beads

For dynamic range reduction in complex biological samples, the potential merits of Equalizer™ beads have been discussed (Thulasiraman *et al.*, 2005, Righetti *et al.*, 2006, section 1.4.2). This was applied to the soluble fraction of chicken skeletal muscle prior to analysis by gel electrophoresis, trypsin digestion and LC-LTQ MSMS to assess the efficiency of this strategy for enrichment of low abundance proteins. Absolute quantification by QconCAT provided the

opportunity to quantify this method of dynamic range reduction, measuring absolute abundance of the proteins incorporated into the QconCAT before and after normalisation.

### 3.4.3 Absolute quantification of the post-translational modification, deamidation using a stable isotope labelled synthetic peptide

The post-translational modification deamidation results in conversion of asparagine to aspartic acid or isoaspartic acid. This is a non-enzymic process (Robinson and Rudd, 1974; www.deamidation.org) and acts to regulate protein degradation (Geiger and Clarke, 1987, Friedman *et al.,* 1991, Deverman *et al.,* 2002, Weintraub and Manson, 2004). Factors influencing the rate of deamidation include temperature, pH and the nature of the flanking amino acids, particularly C-terminal to asparagine (Robinson *et al.,* 2001). Studies using model peptides have determined that a glycine residue in this position achieves the highest rate of deamidation, but it does not follow that this should be true of the intact protein also. Deamidation requires the formation of a cyclic intermediate, for which the peptide backbone and side chain of asparagine must adopt a particular conformation. It is likely that the higher order structure of native proteins will have a greater influence on this flexibility than can be predicted from the amino acid sequence.

Deamidation of a peptide resulting in a mass shift of +0.985Da can be observed in a mass spectrum as a complete or partial reaction from detailed analysis of the peptide isotopomer distribution compared to the expected profile. From such analysis of chicken skeletal muscle soluble proteins digested with trypsin, a peptide from one protein in particular exhibited a noticeable and atypical natural isotope distribution profile, consistent with a mixture of an asparagine containing peptide and the cognate deamidation product. This peptide was derived from the N-terminus of an abundant protein, glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Further analysis confirmed that the 'atypical' isotope profile is attributable to partial deamidation of an asparagine residue and that this is constrained by higher order structure in the native protein. The relationship between release of the N-terminal peptide by proteolysis and subsequent deamidation of the asparagine residue is complex with both reactions exhibiting different rate constants. To construct a comprehensive analysis of the kinetics involved, techniques for absolute quantification using a stable isotope labelled synthetic peptide of identical sequence were applied. Through direct comparison of the rate of deamidation in the model peptide with that observed during proteolysis of the native protein, the absolute effects of higher order structure were elucidated. This illustrates a significant application for absolute

quantification strategies as deamidation of asparagine residues post-proteolysis will influence identification, characterisation and quantification proteomics.

# 4. MATERIALS AND METHODS

## 4. Materials and Generic Methods

Details of methods given in this section are generic and are applied to all appropriate experiments. Further details, specific to each experiment are given in figure legends of Chapter 5: 'Results and Discussion'.

### 4.1 Materials and Reagents

Trypsin (sequence grade) was obtained from Roche Diagnostics (Lewes, UK). All other chemicals and solvents (HPLC grade) were purchased from Sigma-Aldrich Company Ltd (Dorset, UK) and VWR International Laboratory Supplies (Poole, UK).

### 4.2 Preparation and Purification of QconCAT

The artificial QconCAT gene was constructed and synthesised (Beynon *et al.*, 2005; supplementary methods) prior to expression in *E.coli* BL21(λDE3) in mimimal media and stable isotope labelling with $[^{15}N]H_4Cl$ as the sole nitrogen source, or in the presence of $[^{13}C_6]$lysine (100mg/L) and $[^{13}C_6]$arginine (100mg/L) within a mixture of all other amino acids (unlabelled). Expression was induced with isopropyl-β-D-thiogalactopyranoside (IPTG) for up to 6h and the cells were harvested by centrifugation at 8000 x *g* at 4°C for 10 minutes. Inclusion bodies containing QconCAT (as proven by digestion with trypsin and MALDI-ToF MS analysis) were recovered by breaking cells using BugBuster Protein Extraction Reagent (Novagen, Nottingham, UK). Workflow to this point was conducted by Dr. D.M. Simpson and is described in detail (Pratt *et al.*, 2006). Inclusion bodies were resuspended in 20mM phosphate buffer, 6M guanidinium chloride/8M urea, 0.5M NaCl, 20mM imidazole, pH 7.4. From this solution, $[^{15}N]$ labelled, $[^{13}C_6]$lysine/arginine labelled, and unlabelled QconCAT proteins were purified separately by affinity chromatography using a Ni based resin (HisTrap HP Kit, Amersham Biosciences, UK). Following sample loading, HisTrap™ columns were washed with 20mM phosphate buffer, pH 7.4 prior to elution of the sample with the same buffer containing a higher concentration of imidazole (20mM phosphate, 0.5M NaCl, 500mM imidazole, 6M guanidinium chloride/8M urea, pH 7.4) during which phase fractions (1mL) were collected. The purified QconCAT was desalted by three rounds of dialysis against 100 volumes 10mM ammonium bicarbonate, pH 8.5 for 2 h using fresh buffer each time.

### 4.3 Preparation of Chicken Skeletal Muscle Soluble Proteins

Chickens (ISA Brown layer and Ross 308 broiler) were grown to 30d post hatch and several animals of each strain were culled at 1, 3, 5, 10, 20 and 30d at which times, pectoralis muscle was collected (the above procedures were performed at the Roslin Institute, Edinburgh, UK). To isolate the soluble fraction of chicken skeletal muscle, 100mg breast tissue was homogenised in 0.9mL 20mM sodium phosphate buffer, pH7.0 containing protease inhibitors (Complete Protease Inhibitors, Roche, Lewes, UK). This was centrifuged at 15,000 x *g* for 45 minutes at 4°C. The supernatant fraction, containing soluble protein, was then removed. The insoluble fraction was homogenised in the same volume of 20mM sodium phosphate buffer, pH7.0 and centrifugation at 15,000 x *g* for 45minutes at 4°C was repeated; the pooled supernatant fractions (containing soluble protein) were used for all analyses. The total protein concentration of each preparation was measured using a Coomassie Plus Protein Assay (Pierce, Northumberland, UK).

### 4.4 Gel Electrophoresis

#### *One dimensional sodium dodecyl sulphate polyacrylamide gel electrophoresis (1D SDS-PAGE)*

Prior to separation of proteins by molecular weight, samples were heated to 100°C for 5 min with an equal volume of 2X reducing sample buffer (1mL 0.5M Tris buffer, pH 6.8, 1mL glycerol, 0.02g sodium dodecyl sulfate (SDS), 0.01g bromophenol blue, 0.154g dithiothreitol (DTT)). This reduces disulphide bonds (DTT), generates protein:SDS complexes, eliminates higher order protein structure and allows progression of protein migration to be monitored through the gel by adding a blue dye (bromophenol blue). To separate denatured proteins, samples were loaded onto one end of a 12.5% (w/v) polyacrylamide reducing gel and an electric current was applied across the gel causing the negatively charged proteins to migrate towards the cathode (gels were run at 200V for 45 min). Each protein has a constant charge density due to the bound SDS and moves differently according to its size, with larger proteins encountering more resistance through the gel matrix causing them to run more slowly. Following electrophoresis, gels were stained with the dye Coomassie Blue (Bio-Safe:Bio-Rad, Hemel Hempstead, UK) overnight followed by approximately 1h incubation with de-stain solution containing 10% (v/v) acetic acid and 10% (v/v) methanol.

### *Gel image analysis*

Gels were imaged using scanning densitometry using an Epson 160 pro flatbed scanner. For 1D quantification, gel images were converted to black and white TIFF files and band volumes were assessed (Total lab TL100 non-linear, 2006).

## 4.5 PROTEOLYSIS

### *Proteolysis for protein identification and quantification*

Proteins were digested to peptides with the protease trypsin; proteins were diluted in ammonium bicarbonate (50mM, pH8.8) and incubated with trypsin at a ratio of 100:1-10:1 (protein:trypsin). For identification, proteins were digested following separation by 1D SDS-PAGE; a gel plug or slice containing protein material was excised and de-stained using 50:50 acetonitile:50mM ammonium bicarbonate, dehydrated with acetonitrile and digested overnight with trypsin.

### *Proteolysis kinetics*

To investigate the kinetics of proteolysis for individual peptides, stable isotope labelled internal standard peptides were added to the digestion reaction in known amounts. This provided a reference standard upon which signal intensity of peptides cleaved from intact proteins was reconciled to obtain the rate of digestion. For this, proteins were digested with trypsin and samples of this mixture were removed at selected time points where the reaction was stopped by addition to a strong acid (10% (v/v) formic acid). Disappearance of intact proteins and appearance of limit peptides was monitored by 1D SDS-PAGE and MALDI-ToF MS. This method was also used to investigate the effect of denaturing protein structure prior to addition of enzyme on proteolysis kinetics of analyte proteins in solution.

## 4.6 MASS SPECTROMETRY

Two mass spectrometers were used to acquire quantitative data for the research reported in this thesis, MALDI-ToF MS (M@LDI, Waters, Manchester, UK) and ESI-Q-ToF MS (Waters, Manchester, UK) with some supplementary data obtained using a different MALDI-ToF MS instrument (AXIMA MALDI ToF$^2$, Shimadzu, Manchester, UK). In addition, protein identification data was acquired using a linear quadrupole ion trap (LTQ, Thermo Scientific, Hemel Hempstead, UK). Chromatography platforms are detailed in section 4.7.

***MALDI-ToF MS***

The matrix solution (alpha-cyano hydroxycinnamic acid, 10mg/mL in 0.1% (v/v) TFA, 50% (v/v) acetonitrile) which absorbs energy at the wavelength of a UV laser irradiating the sample plate was spotted over the sample deposited onto a stainless steel target. Upon laser irradiation, peptides enter the gas phase, absorbing energy and becoming protonated; predominantly to singly charged ions $[M+H]^+$. Ions are accelerated from the source into the flight tube with kinetic energy directly related to their mass and velocity ($KE=\frac{1}{2}mv^2$) with smaller ions travelling along the flight tube faster than larger ions with less kinetic energy. To improve the resolution of the resulting mass spectrum, MALDI-ToF (M@LDI; Waters, Manchester, UK) has a lengthened flight tube via the use of a reflectron, or 'ion mirror' which reflects the ions off axis to the detector. This focuses the ions, compensating for small discrepancies in kinetic energy as ions of the same m/z with greater kinetic energy penetrate the reflectron further and consequently leave the reflectron at the same time as those with less kinetic energy (of the same m/z). Resolution can also be improved using MALDI and ToF MS by delaying the pulse of ion extraction, focusing the ions immediately following ionisation. This generates ions with a significantly smaller kinetic energy distribution so that all ions of the same m/z enter the flight tube at the same time, and with the same kinetic energy. For analysis by MALDI-ToF MS, peptide mixtures (1µL) were mixed with an equal volume of α-cyano-hydroxycinnamic acid in 50% (v/v) acetonitrile, 0.1% (v/v) trifluoroacetic acid on a stainless steel 96 well MALDI target and allowed to air dry. A four point calibration of known peptides was carried out to test the sensitivity and mass accuracy of the instrument prior to peptide detection over a range of 900-3000m/z. For each combined spectrum, 20-30 spectra were acquired (laser energy typically 30%) with 10 shots per spectrum and a laser firing rate of 5Hz. Data were processed using MassLynx software to subtract background noise using polynomial order 10 with 40% of the data points below this polynomial curve and a tolerance of 0.01. Spectral data were also smoothed by performing two mean smooth operations with a window of three channels. For MALDI-ToF MS (AXIMA MALDI ToF$^2$, Shimadzu, Manchester, UK), peptides were analysed with the same sample preparation and data acquisition. Spectra were processed using an average smooth filter of width two channels and a baseline subtract filter width of 15 channels.

***ESI Q-ToF MS/MSMS***

Sample containing peptides or proteins is sprayed from a high voltage needle into the electrospray source which is maintained at a constant potential difference across the sample cone. Solvent particles containing peptides are forced into the gas phase in the source and peptides become protonated; gaining multiple positive charges. Tryptic peptides are typically doubly charged $[M+2H]^{2+}$ although longer peptides and those with other basic residues (for example histidine) that can be protonated at additional sites have a higher charge state. Charge state is easily determined from the m/z difference between the monoisotopic and the first [$^{13}C$] peak, for example a difference of 0.5m/z units denotes a charge state of two. Ions are selected by mass to enter the quadrupole by adjusting the potential across it. The trajectory of the ions, as a function of time and position of the ion from the centre of the rods, providing the ion is stable in the quadrupole is measured to separate the ions according to their mass to charge ratio (m/z). Ions then travel directly to the flight tube and are detected. For tandem mass spectrometry, specific ions are selected to pass through the quadrupole, and are collided with an inert gas such as argon or helium in a collision cell causing them to fragment. Fragment ions pass through to the flight tube and are detected by their m/z. Peptides were acquired over the range 400-2000m/z with the capillary voltage set at 1900V, collision energy 10V and sample cone at 55V for LC-MS analysis. For LC-MSMS using ESI Q-ToF MS, collision energy was increased to 30%.

ESI Q-ToF MS (Waters, Manchester, UK) was also used to analyse intact proteins under the same instrument conditions, acquiring over the range 700-1800m/z. Mass spectra produced were highly complex with multiple overlapping charge envelopes for each protein. This was resolved using deconvolution software, for example MaxENT1 maximum entropy software in MassLynx to produce a true molecular mass spectrum.

***Quadrupole ion trap MSMS***

A linear quadrupole ion trap (LTQ, Thermo Scientific, Hemel Hempstead, UK) was used for high throughput protein identification by tandem mass spectrometry. Peptides ionised by ESI enter the trap where a combination of radio frequency voltages are applied to the rods and a direct current is applied to the end lenses at both sides of the trap to destabilise successive ion trajectories, thus expelling ions of a selected mass. For fragmentation, selected precursor ions are retained inside the trap where collision energy is applied through an inert gas, for example helium, causing the peptide ion to fragment. Multiple stages of mass spectrometry by

subsequent isolation and fragmentation of product ions from the first fragmentation may be performed. For protein identification, tryptic peptides were ionised by electrospray, and ions were determined over the range 400-1500m/z with the capillary voltage at 50V, spray voltage at 1.8kV.

## 4.7 Liquid Chromatography

### *Reversed-phase high performance liquid chromatography*

Peptides are separated in solution on the basis of their hydrophobicity prior to mass spectrometric analysis. This is performed in a column packed with silica based beads with surface bound long n-alkyl groups for example n-octadecyl ($C_{18}$), covalently bound. Peptides bind to the matrix and are eluted using a gradient of an organic solution, for example acetonitrile, with the most hydrophobic peptides eluting at the end of the gradient in a high concentration of organic solvent. For quantification using ESI Q-ToF MS and MALDI-ToF MS, peptides were separated using an EASY-nLC (Proxeon, Denmark) nanoflow system. Nanoflow HPLC at 200nL/min was used to resolve peptides (in 0.1% (v/v) formic acid) over a 50 minute acetonitrile gradient (0-100%). For ESI Q-ToF MS, peptides were eluted directly from the analytical column and infused into the source whereas for MALDI-ToF MS, peptides were eluted from the column and manually spotted onto a MALDI target at selected time intervals, allowed to dry and covered with matrix. For LTQ-MSMS analysis, peptides from in-solution or in-gel digests were separated using an Ultimate 3000 HPLC system (Dionex, UK). Nanoflow HPLC at 300nL/min was used to resolve peptides (in 0.1% (v/v) formic acid) over a 60 minute acetonitrile gradient (0-100%).

## 4.8 Protein Identification

### *Peptide mass fingerprinting*

Analyte proteins were digested in-gel or in-solution with trypsin (as described in section 4.5). Peptides were analysed by MALDI-ToF MS and monoisotopic masses were entered into the MASCOT search engine. Data were searched against the database MSDB for taxonomy: Chordata, variable modifications: oxidation of methionine, protease: trypsin, missed cleavages: 1, peptide tolerance: 250ppm. MOWSE scores above 65 at probability level, p=0.05 were accepted as confident matches. Peptide digestion maps (Beynon, 2005) were created indicating sequence coverage, including peptides that were identified as part of a missed cleavage.

***MSMS ion search***

Proteins were separated by 1D SDS-PAGE and digested in-gel with trypsin prior to analysis of peptides by LC-ESI-LTQ MSMS. MSMS data were searched against MSDB using MASCOT with the following parameters; taxonomy: Chordata, protease: trypsin, variable modifications: oxidation of methionine, peptide tolerance: 250ppm, MSMS tolerance: 250ppm, peptide charge: 1+, 2+ and 3+, instrument: ESI-TRAP, from which only confident identifications (MOWSE score>45, $p<0.05$) were accepted.

## 4.9 Protein Quantification

Analyte proteins were mixed with QconCAT, or stable isotope labelled internal standard peptides and digested with trypsin. Peptides were analysed by MALDI-ToF MS or ESI-Q-ToF MS/MSMS from which relative signal intensity of analyte:internal standard was used for quantification. For quantification of chicken skeletal muscle soluble proteins using the QconCAT method, QconCAT (7µg) and chicken skeletal muscle soluble proteins (70µg) from broiler and layer chickens 1d-30d post-hatch (6 time points, 4 birds at each time point), were mixed and diluted 10 fold with 50mM ammonium bicarbonate, and 10% (v/v) acetonitrile, prior to addition of trypsin (20:1 substrate:protease). The reaction mixture was incubated at 37°C for 24h after which the digest was incubated with additional trypsin (20:1 substrate:protease) to achieve complete digestion and 1µL was analysed by MALDI-ToF MS.

## 4.10 Normalisation of Chicken Skeletal Muscle Soluble Protein Abundance using Equalizer™ Beads

20mg Prospectrum-2 (Louisville, KY, USA) beads were washed in 1mL 50% (v/v) MeOH and mixed gently for 10min. Beads were allowed to settle and the supernatant was removed and discarded. MeOH 50% (v/v) was added to cover the surface of the beads that were left to swell overnight at 4°C. Once swollen, 20mg beads (constituting 100µL settled bed volume) were transferred to a 1.5mL Eppendorf tube. Beads were washed in 1mL double distilled $H_2O$ in a roller mixer for 30min prior to equilibration by repeated washing in 20mM sodium phosphate buffer pH7.0 for 30min. After each wash, beads were left to settle for 5min and the supernatant was removed. Approximately 1mL sample containing 25mg, 50mg and 100mg soluble protein in three separate experiments was added to the beads and mixed for 2h on a roller mixer. Unbound protein was collected as the supernatant fraction after beads had settled for 5min.

The beads were subsequently washed eight times in 1mL 20mM phosphate buffer and supernatant fractions were removed and collected.

## 5. Results and Discussion


5.1 Design, preparation, purification and analysis of QconCAT and analyte proteins ..... 58
5.1.1 Design, preparation and purification of QconCAT ..... 58
5.1.2 Proteomic analysis of QconCAT ..... 59
5.1.3 Proteomic analysis of chicken skeletal muscle soluble proteins ..... 63

5.2 Proteolysis of QconCAT and analyte proteins ..... 63
5.2.1 Proteolysis of QconCAT ..... 63
5.2.2 Proteolysis of chicken skeletal muscle soluble proteins ..... 65

5.3 Sample complexity and dynamic range ..... 67
5.3.1 Mass spectrometry for absolute quantification using the QconCAT method ..... 67
5.3.2 Challenges for data acquisition and analysis for quantification ..... 68

5.4 Validation of the QconCAT method ..... 71
5.4.1 Quantification of unlabelled QconCAT by labelled QconCAT ..... 71
5.4.2 Variance in the QconCAT method ..... 71
5.4.3 Accuracy of the QconCAT method ..... 72
5.4.4 Comparison of the QconCAT method with alternative strategies for absolute quantification ..... 72

5.5 Absolute quantification of chicken skeletal muscle soluble proteins ..... 75

5.6 Additional applications of QconCAT technology ..... 78
5.6.1 Quantification of soluble skeletal muscle proteins in other species ..... 78
5.6.2 Quantification of normalisation using Equalizer™ bead technology ..... 79
5.6.3 Absolute quantification of the post-translational modification, deamidation ..... 82

# 5. Results and Discussion

## 5.1 Design, Preparation, Purification and Analysis of QconCAT and Analyte Proteins

### 5.1.1 Design, preparation and purification of QconCAT

To measure the absolute amount of chicken skeletal muscle soluble proteins during growth of chickens bred for meat (broiler) and those bred for eggs (layer), a group of twenty soluble proteins was selected to be quantified using a single QconCAT. For each of these, a representative peptide was chosen that gave a strong signal in previous MALDI-ToF MS analyses of tryptic digests (Beynon *et al.*, 2005; Table 3, section 1.5.2). The peptides were used to guide construction of the DNA sequence of the QconCAT, which was synthesised, inserted into a pET21a vector and expressed in *E.coli* grown in labelled ($[^{15}N]H_4Cl$) or unlabelled ($[^{14}N]H_4Cl$) media (Pratt *et al.*, 2006). For QconCAT expression, a typical bacterial culture of 200mL was induced at an $OD_{600}$ of 0.6-0.8 which generated 5-10mg of QconCAT after cell breakage, recovery of inclusion bodies and affinity chromatography of 8M urea solubilised protein on 1mL NiNTA columns. After induction, the QconCAT protein was visible as a major band in 1D SDS-PAGE of a broken cell preparation (work flow to this point conducted by Dr. D.M. Simpson) and analysis of QconCAT purification fractions by 1D SDS-PAGE (Figure 16) revealed a major band at approximately 35kDa as expected from the predicted mass of the protein from its sequence. For purification, inclusion bodies containing expressed QconCAT protein were solubilised in 8M urea and purified by affinity chromatography using a Ni based resin (HisTrap HP Lit, Amersham Biosciences, UK). Following sample loading, columns were washed and protein was eluted with 500mM imidazole into five 1mL fractions. For the [$^{15}N$] labelled QconCAT, the protein was contained predominantly in fractions 1-3 and unlabelled QconCAT in fractions 1 and 2. Presence of other bands immediately above the main QconCAT band were presumed to be modified products of the same protein and will be discussed in more detail in section 3.1.1. A small amount of protein material was also eluted in the wash and flow through fractions. Protein containing fractions following affinity purification were pooled and the protein concentration was determined using a Coomassie-Plus protein assay (Pierce, Northumberland, UK). Each fraction was aliquoted to 50µL and stored at -20°C.

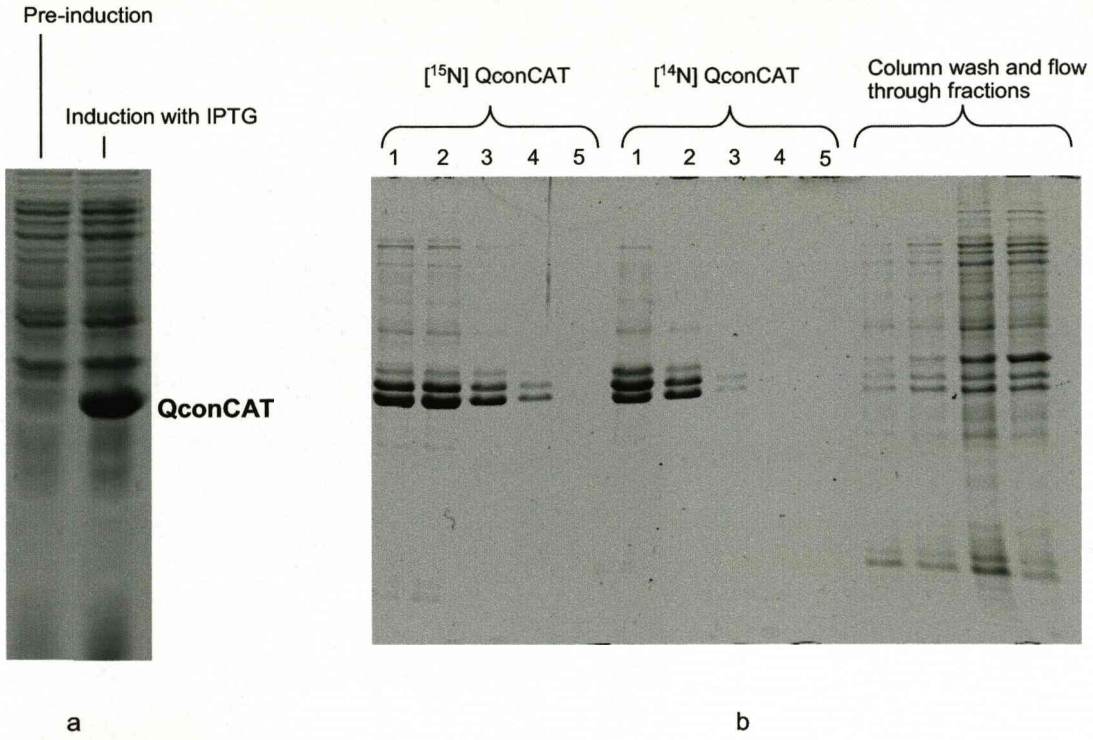



**Figure 16. QconCAT expression and purification.**
The artificial QconCAT gene (Beynon *et al.*, 2005) was expressed in *E.coli* in minimal medium containing $[^{15}N]H_4Cl$ as the sole nitrogen source. Expression was induced with isopropyl-β-D-thiogalactopyranoside (IPTG) and the cells were harvested by centrifugation at 1400 x g at 4°C for 15 minutes. Inclusion bodies containing QconCAT were recovered by breaking cells using BugBuster Protein Extraction Reagent (Novagen, Nottingham, UK). Expression and analysis by 1D SDS-PAGE before and after induction (a) was completed by Dr. D.M. Simpson (Proteomics and Functional Genomics, University of Liverpool). Inclusion bodies were re-suspended in 20mM phosphate buffer, 8M urea, 0.5M NaCl, 20mM imidazole, pH 7.4. From this solution, labelled ($^{15}N$) and unlabelled QconCAT proteins were purified separately by affinity chromatography using a Ni based resin (HisTrap HP Kit, Amersham Biosciences, UK). Following sample loading, HisTrap™ columns were washed with 20mM phosphate buffer, pH 7.4 prior to elution of the sample with the same buffer containing a higher concentration of imidazole (20mM phosphate, 0.5M NaCl, 500mM imidazole, 8M urea, pH 7.4) during which phase five fractions (1mL) were collected and analysed by 1D SDS-PAGE (b). The first five lanes contain 5µL of each 1mL fraction of purified $[^{15}N]$ QconCAT (labelled/heavy; H), the next five contain 5µL of each 1mL fraction of purified $[^{14}N]$ QconCAT (unlabelled/light; L). Column washes with binding buffer containing 8M urea and column flow through collected when QconCAT protein was loaded onto HisTrap™ columns were also analysed and are shown as the last four lanes.

### 5.1.2 Proteomic Analysis of QconCAT

The intact mass of the protein measured by ESI-Q-ToF MS was 33036Da (Figure 17), this is consistent with the predicted mass of 33167Da for the QconCAT protein with the loss of the initiator methionine residue from the N terminus (131Da) by the action of methionine aminopeptidase following translation and sufficient synthesis of the protein (Ben-Bassat *et al.*, 1987). For *E.coli* proteins, the first amino acid is always N-formylmethionine which results from a post acylation modification of the methionine on a specific tRNA. The formyl group is removed by the action of a specific formylase yielding an unmodified N-terminal methionine which is subsequently removed (Adams and Capecchi, 1966, Adams, 1968). Some adducts were observed in the intact mass analysis; the three most intense peaks of greater mass than the true QconCAT protein were attributed to addition of one, two and three sodium (Na) groups and are presumed to result from salt in the sample preparation, for example NaCl in the binding buffer for affinity purification.

Purified, unlabelled QconCAT was subjected to in-solution digestion with trypsin prior to analysis of peptides by MALDI-ToF MS (Figure 18). All of the predicted QconCAT peptides within the mass range 900-3000m/z were observed, although not of equal intensity, despite being present in equal amounts. The influences and effects of ionisation inherent with MALDI-ToF MS analysis are crucial for absolute quantification and are discussed in section 3.3. All ions in the peptide mass fingerprint of unlabelled QconCAT protein were accounted for except peaks at 1275.80m/z and 1233.80m/z, each of which was 17m/z less than the genuine tryptic peptides T8 (QVVESAYEVIR) and T18 (QVVESATEVIK), both representing isoforms of lactate dehydrogenase. As both of these peptides are of identical sequence apart from the C-terminal residue, the most likely event was post-proteolytic modification of the N-terminal glutamic acid which had cyclised to form pyroglutamic acid. To confirm this, peptides were analysed by ESI-Q-ToF MSMS and the doubly charged peak at 638.4m/z ($[M+H]^+$ 1275.80m/z) was fragmented by MSMS and sequenced *de novo*. The majority of the peptide sequence was allocated from y-ions, confirming the identity of this peptide (Figure 19). [$^{15}$N] labelled QconCAT was also digested in-solution with trypsin and analysed by MALDI-ToF MS, allowing comparison of labelled and unlabelled mass spectra (Figure 20). [$^{15}$N] labelled; 'heavy' peptides were distinguished from unlabelled; 'light' peptides by a mass shift dependent on the number of nitrogen (N) atoms in each peptide. For example, T12 of sequence VLYPNDNFFEGK, $[M+H]^+$ 1442.82m/z containing 15N atoms is increased in mass by 15Da due to the 'heavy' label.

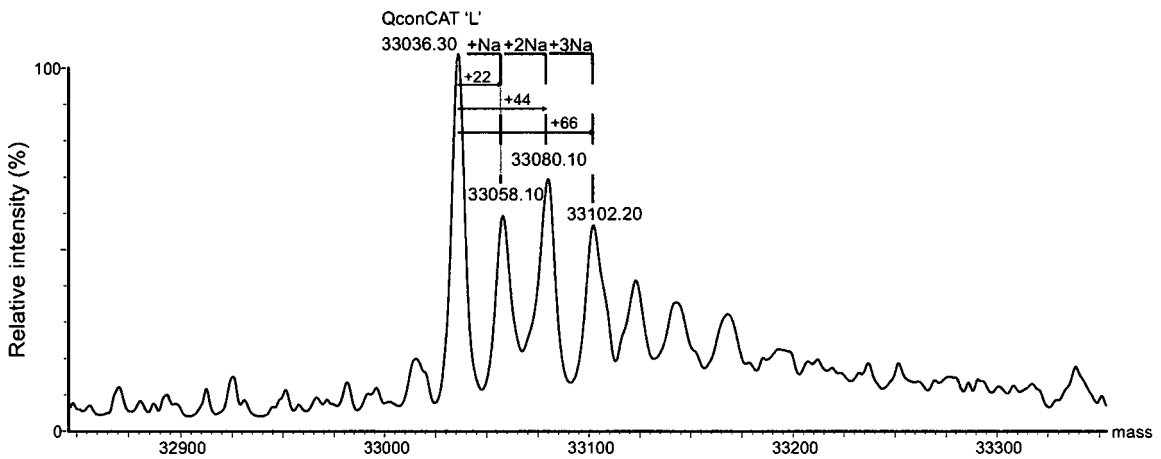



**Figure 17. ESI-Q-ToF MS analysis of unlabelled QconCAT protein.**
QconCAT protein, solubilised and purified in 8M urea, was diluted 100 fold with 1% (v/v) formic acid and the intact mass was determined by ESI-Q-ToF MS by direct infusion into a Waters QToF (Waters, Manchester, UK). Mass spectra were acquired over the m/z range 700-1800 at 10% collision energy with sample cone voltage 55V and capillary voltage 1900V. The combined mass spectrum (100 scans) was deconvoluted using the MaxENT 1 maximum entropy algorithm (MassLynx) between 32800 and 33400Da at a resolution of 0.1Da/channel to recover the true mass of the protein.

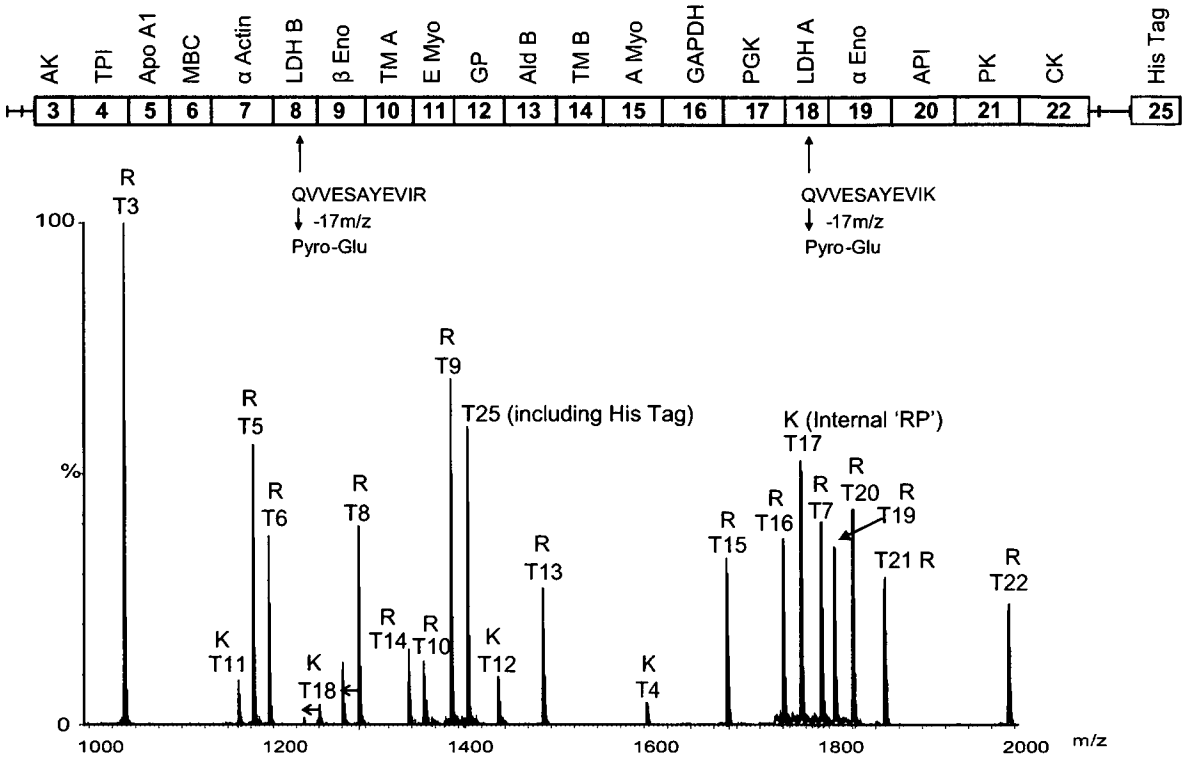



**Figure 18. QconCAT protein digested in-solution with trypsin.**
The QconCAT protein was purified and digested in-solution. For this, the QconCAT protein was diluted to 5µM in 50mM ammonium bicarbonate and digested with trypsin (20:1 substrate:protease) at 37°C for 24h. Peptides were analysed by MALDI-ToF MS (M@LDI; Waters, Manchester, UK). Each peak in the spectrum is the surrogate peptide for a different protein as indicated by the peptide map. Peptide ions resulting from chemical modification of N-terminal glutamine residues have also been indicated, thus explaining every major ion in the spectrum. The protein names, and their abbreviations are: AK: adenylate kinase, TPI: triose phosphate isomerase, ApoA1: apolipoprotein A1, α Actin alpha actin, α Eno: alpha enolase, β Eno: beta enolase, E Myo: embryonic myosin, A Myo: adult myosin, GAPDH: glyceraldehyde 3-phosphate dehydrogenase, API: actin polymerisation inhibitor, PK: pyruvate kinase, CK: creatine kinase, LDH A: lactate deydrogenase A, LDH B: lactate dehydrogenase B, GP: glycogen phosphorylase, TM A: tropomyosin A and TM B: tropomyosin B.

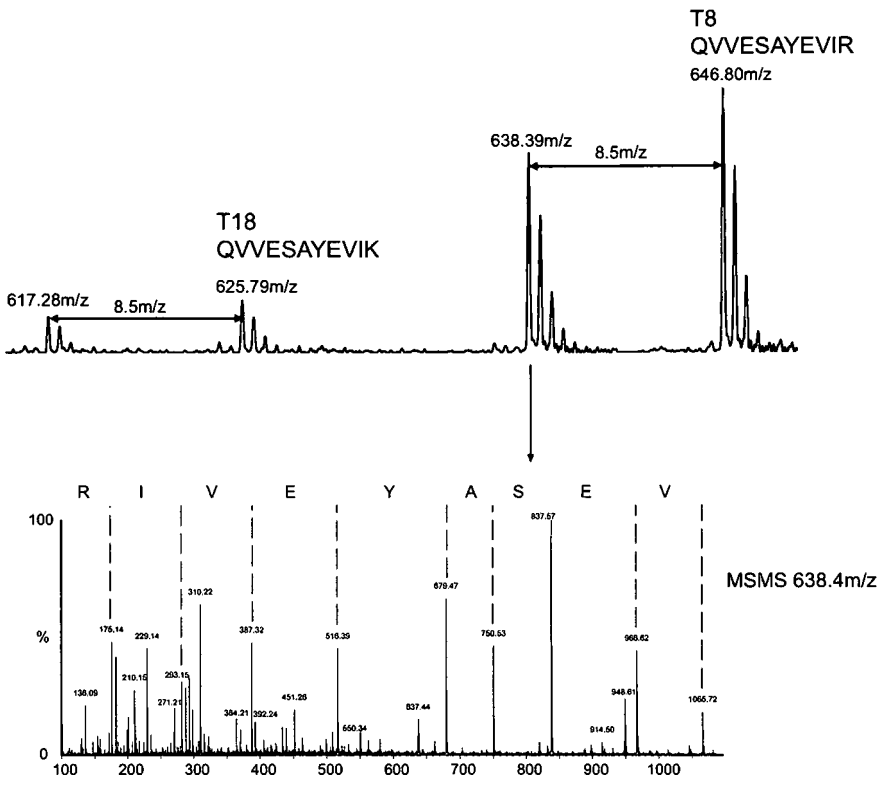



**Figure 19. Confirmation of pyro-glutamic acid modification to QconCAT peptides.**
QconCAT protein was diluted to 5µM in 50mM ammonium bicarbonate and digested with trypsin (20:1 substrate:protease) at 37°C for 24h. Peptides were analysed by ESI-Q-ToF MS (Waters, Manchester, UK). Doubly charged peptides T8 and T18 representing two isoforms of lactate dehydrogenase that differ only in the C-terminal amino acid both gave rise to an ion less 17Da in mass. Tandem mass spectrometry was performed on the doubly charged peptide at 638.4m/z using 30% collision energy.

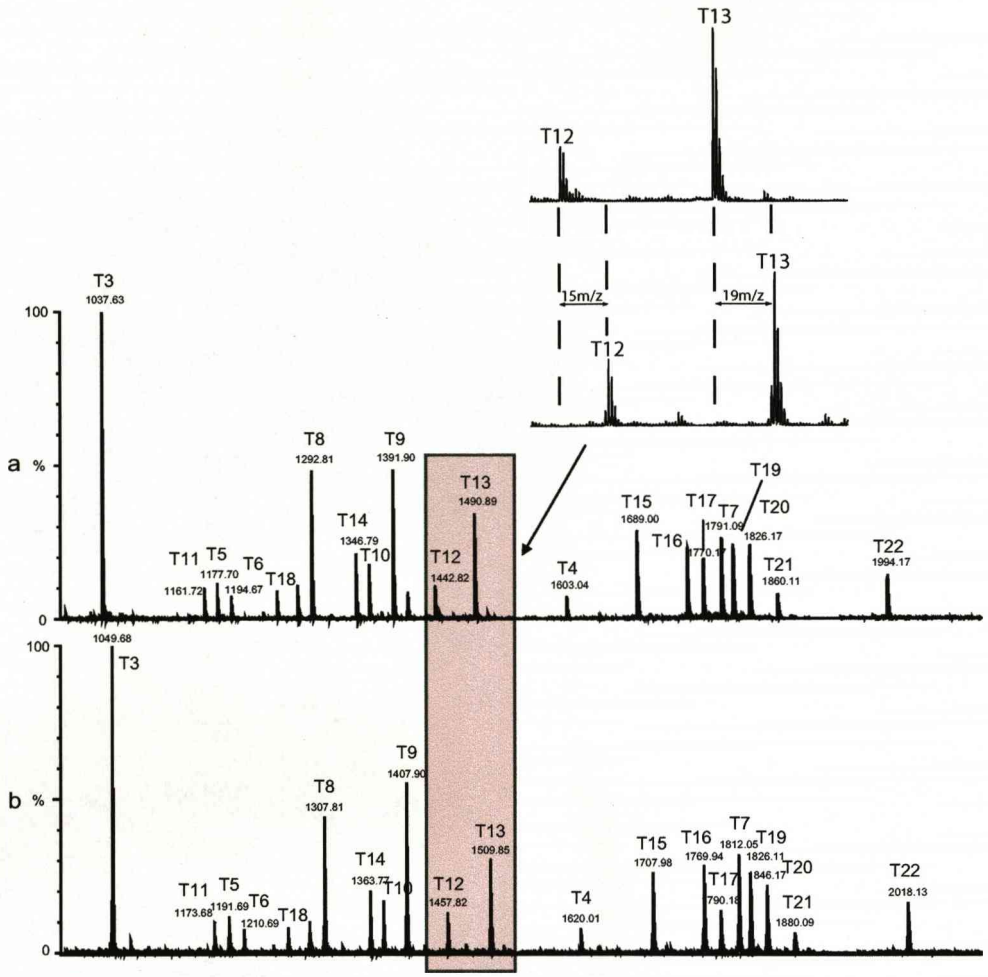



**Figure 20. Distinction between unlabelled and [$^{15}$N] labelled QconCAT peptides in MALDI-ToF mass spectra.**

Unlabelled (a) and [$^{15}$N] labelled (b) QconCAT proteins were diluted to 5µM in 50mM ammonium bicarbonate and digested with trypsin (20:1 substrate:protease) at 37°C for 24h. Peptides were analysed by MALDI-ToF MS. Labelled peptides are 'heavier' in mass than their unlabelled counterparts according to the number of nitrogen atoms in the peptide. A region of the spectrum containing two Q-peptides, T12 and T13 is highlighted and this region is expanded and inserted above the main spectrum, indicating the mass offset between 'heavy' and 'light' peptides.

Analysis of QconCAT preparations by 1D SDS-PAGE revealed two major bands, and several protein bands in less abundance in the same region as the QconCAT protein. This phenomenon was particularly pronounced for samples that had been stored for a considerable amount of time in binding buffer containing 8M urea at -20°C, or if freeze thawed several times prior to analysis (results not shown). In-gel digestion of the two main bands with trypsin and MALDI-ToF MS analysis of the peptides revealed no significant differences in the spectra (Figure 21), although there was limited evidence of missed cleavage peptides in the lower of the two bands. It was expected that the presence of higher molecular weight bands would be caused by aggregation of the protein product, most likely associated with the formation of sample contaminant adducts, for example with salts. The mass range of MALDI-ToF MS is not sufficient to analyse such products upon trypsin digestion at high resolution and in any case, the efficiency and accessibility of trypsin would need to be carefully controlled to ensure these are not artifacts of insufficient cleavage other than that resulting from structural impediments. However, upon in-solution QconCAT digestion with trypsin and subsequent MALDI-ToF MS analysis of peptides, additional ions mass shifted by 43Da from each tryptic peptide were apparent. This is consistent with carbamylation of peptides (Figure 22) occurring as the result of isocyanic acid formation from urea in equilibrium with ammonium cyanate in solution. Isocyanic acid reacts with free amino groups, for example the N-terminus and lysine side chains and may compromise quantification, depending on the extent of carbamylation. It was therefore essential to remove urea prior to analysis of QconCAT, or to refine the purification method so as to eliminate the use of urea in binding and elution buffers. De-salting was carried out prior to MS analysis by filtration using Sephadex G25 spun columns. However, this was associated with a considerable degree of protein loss and an alternative, $C_4$ MicroTrap™ (Presearch, Basingstoke, UK) columns were investigated (Figure 23). This method was effective for removing the effects of salts on the QconCAT protein, although evidence of other bands in the same region can still be seen on the gel. For quantification, the preparation of QconCAT protein must be homogeneous, and consequently other methods were investigated to improve purification. These included eluting QconCAT from HisTrap™ columns without urea in the buffer. However, the inclusion bodies produced from the expression strain required solubilisation, thus QconCAT material was loaded onto HisTrap™ columns within an 8M urea solution. In order to elute without urea once the protein had been loaded, the columns were first washed with binding buffer containing urea, and then without, before eluting the protein from the column in elution buffer without urea. This was successful in producing a homogeneous preparation of QconCAT protein (Figure 24). However, since the resulting QconCAT solution

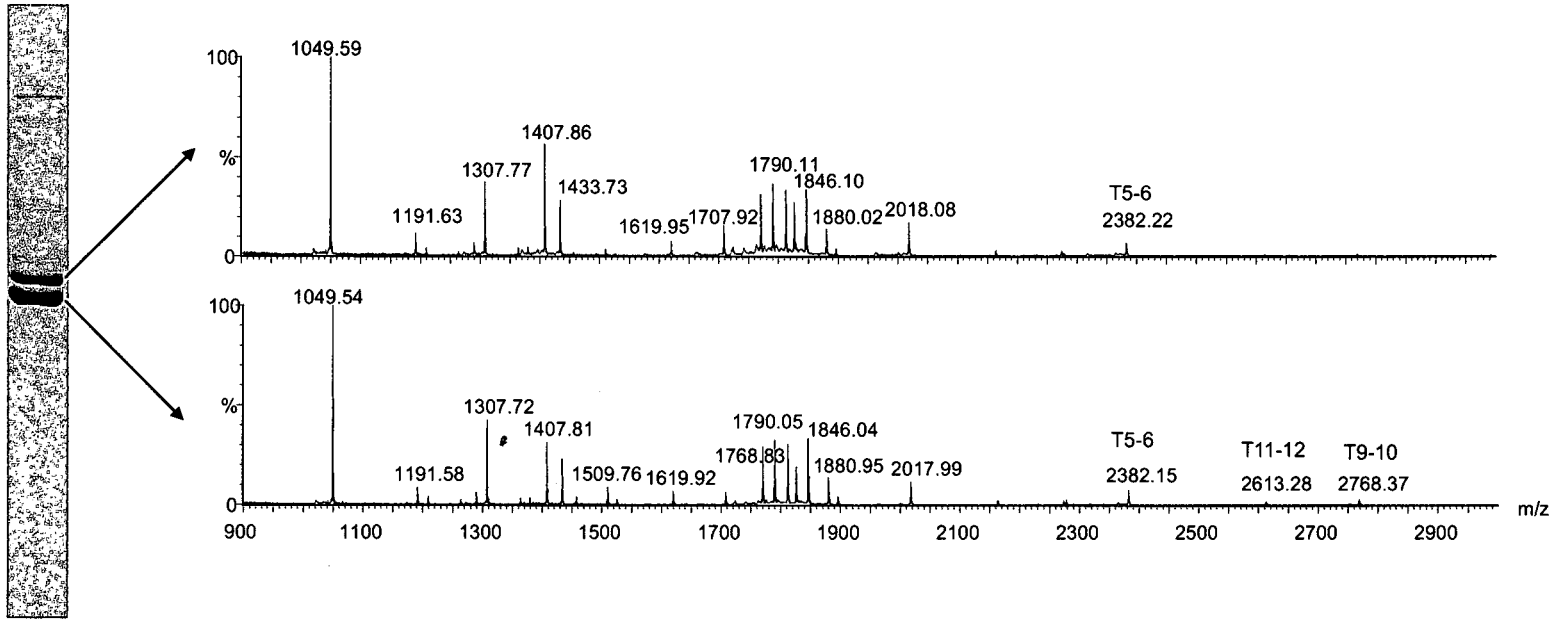
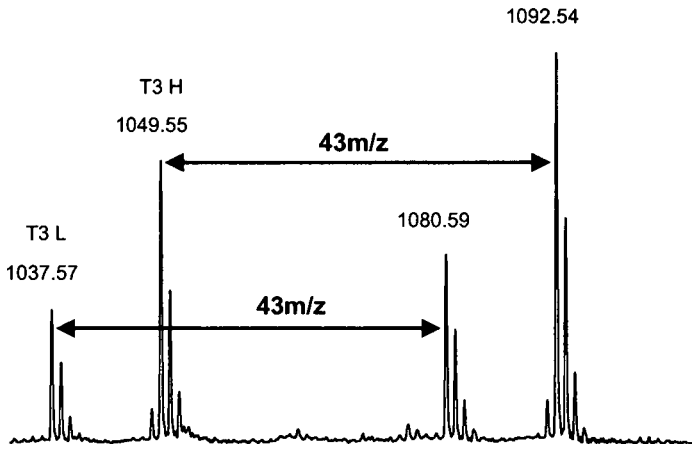
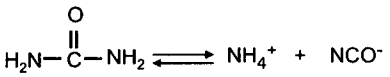



**Figure 21. Diagnostic peptide mass fingerprinting to distinguish between two major protein bands upon SDS-PAGE separation of QconCAT protein.** QconCAT protein was solubilised and purified in 8M urea prior to analysis by SDS-PAGE. The two major bands were excised, de-stained using 50:50 acetonitrile:50mM ammonium bicarbonate, dehydrated with acetonitrile and digested overnight in-gel with trypsin. Peptides were analysed by MALDI-ToF MS.

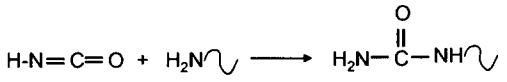



Decomposition of urea



Urea Ammonium cyanate

Carbamylation of peptides and proteins



Isocyanic acid | Peptide amino terminus | Carbamylated peptide or protein

**Figure 22. Carbamylation of QconCAT peptides.**
Unlabelled and labelled QconCAT proteins were mixed, diluted to 5µM in 50mM ammonium bicarbonate and digested with trypsin (20:1 substrate:protease) at 37°C for 24h and peptides were analysed by MALDI-ToF MS. Additional ions present at 43m/z from Q-peptides were attributed to decomposition of urea and subsequent carbamylation of peptides.

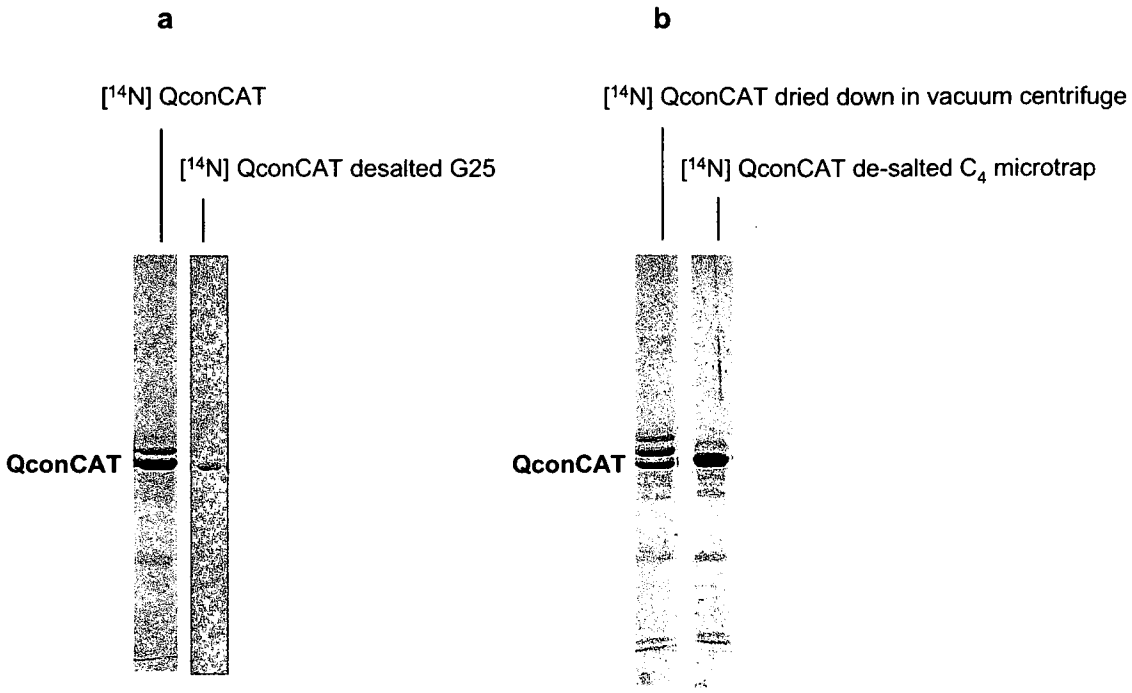



**Figure 23. Implementation of strategies to de-salt QconCAT protein preparations.**

Unlabelled QconCAT purified fractions in 8M urea de-salted with a) sephadex G25 spun columns and b) $C_4$ MicroTraps™. For sephadex G25 spun columns, 250µL sephadex in a 500µL eppendorf tube was washed with $ddH_2O$ and spun at 2000 rpm for 2 minutes when sample was added. The sample was recovered in the original volume. For $C_4$ Microtrap™ columns, these were washed in 0.1% (v/v) formic acid and the sample was eluted in 90% (v/v) acetonitrile/ 0.1% (v/v) formic acid. For SDS-PAGE, samples were dried down in a vacuum centrifuge and reconstituted in reducing sample buffer containing SDS.

did not contain a solubilising agent, the protein was observed to precipitate out of solution both relatively rapidly following purification and upon freeze-thawing. To avoid this, the alternative solubilising agent guanidine hydrochloride (GHCl) was used and QconCAT was solubilised in the same way, replacing 8M urea in all buffers with 6M GHCl during purification. Purified protein was eluted from the column in 1mL fractions which were analysed by 1D SDS-PAGE (Figure 25). As GHCl reacts with SDS in the sample buffer forming a guanidinium dodecyl sulphate complex which precipitates out of solution, GHCl was not compatible with SDS-PAGE, and was removed from the sample prior to downstream analysis. This was achieved using dialysis against 1mM ammonium bicarbonate and will be discussed in the context of the next iteration of QconCAT expression and purification.

The first QconCAT protein designed for chicken skeletal muscle was labelled by growing the expression strain in [$^{15}$N]$H_4Cl$ as the sole nitrogen source. Thus each peptide incorporated [$^{15}$N] instead of [$^{14}$N] with the mass offset of each labelled peptide determined by the number of nitrogen atoms in the peptide. Although this provided an efficient labelling strategy, [$^{15}$N] is a non-uniform label for peptides and proteins as the mass difference between each analyte and internal standard peptide is sequence dependent and varies for each peptide, making it more difficult to distinguish peptide pairs in complex mass spectra. In addition, the relatively high natural abundance of [$^{14}$N] can lead to incomplete labelling. An alternative strategy is to label specific amino acids, for example lysine and arginine, with [$^{15}$N] or [$^{13}$C]. As both of these amino acids contain six carbon atoms, and each tryptic peptide contains only one of these amino acids, [$^{13}$C] labelling of arginine and lysine provides a uniform strategy where the mass offset between unlabelled and labelled peptides is a constant 6Da. Stable isotope labels can also be used in combination to label specific amino acids, thus providing a specific mass difference between 'light' and 'heavy' peptides, for example [$^{13}C_6$][$^{15}N_4$]-arginine which is 'heavier' by 10Da than [$^{12}C_6$][$^{15}N_4$]. For tryptic peptides, the exception to the uniform incorporation of stable isotope label is the presence of a proline residue next to the cleavage site for trypsin, as in tryptic Q-peptide T17 representing the protein phosphoglycerate kinase. Here the peptide contains an internal arginine and a C-terminal lysine residue, thus the mass offset is 12Da. The QconCAT protein was fully labelled with [$^{13}C_6$]arg/[$^{13}C_6$]lys by growing the expression strain in media with all essential amino acids added unlabelled except for lysine and arginine which are added labelled with [$^{13}$C] (labelling carried out by Dr D.M. Simpson). During protein synthesis, each lysine and arginine residue was incorporated into the protein in labelled form. QconCAT protein both labelled and unlabelled was expressed, solubilised and purified with 8M urea and

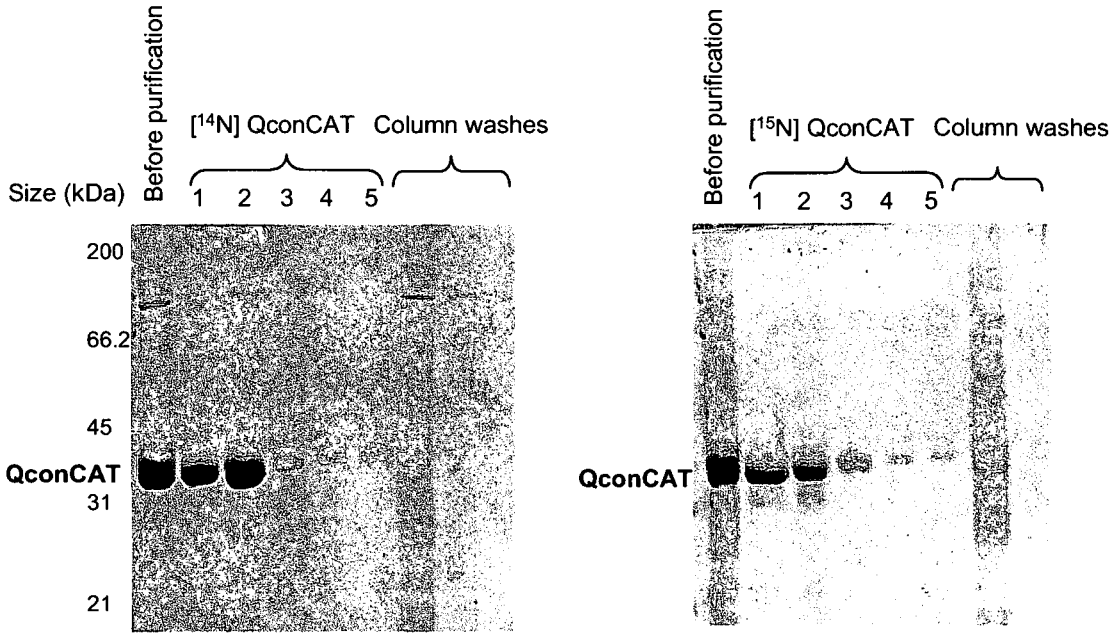



**Figure 24. Purification of QconCAT labelled (H) and unlabelled (L) proteins using HisTrap™ columns without prior solubilisation in 8M urea.**

QconCAT protein was solubilised in 20mM phosphate buffer, 8M urea, 0.5M NaCl, 20mM imidazole, pH 7.4. From this solution, labelled [$^{15}$N] and unlabelled [$^{14}$N] QconCAT proteins were purified separately by affinity chromatography using a Ni based resin (HisTrap HP Kit, Amersham Biosciences, UK). Following sample loading, HisTrap™ columns were washed with 20mM phosphate buffer, pH 7.4 prior to elution of the sample with the same buffer containing a higher concentration of imidazole and no urea (20mM phosphate, 0.5M NaCl, 500mM imidazole, pH 7.4) during which phase five fractions (1mL) were collected and analysed by 1D SDS-PAGE.

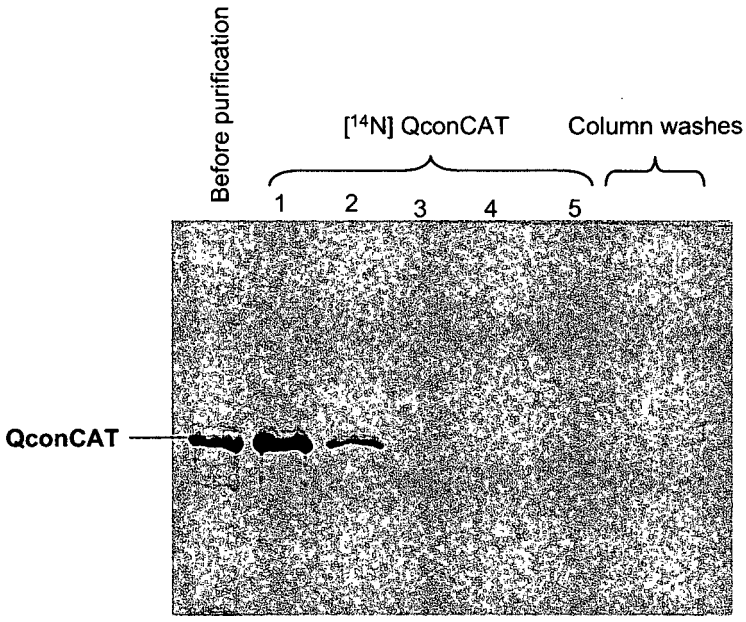



**Figure 25. Purification of unlabelled QconCAT protein, solubilised in 6M guanidinium chloride.**

QconCAT protein was solubilised in 20mM phosphate buffer, 6M guanidinium chloride, 0.5M NaCl, 20mM imidazole, pH 7.4. From this solution, QconCAT protein was purified by affinity chromatography using a Ni based resin (HisTrap HP Kit, Amersham Biosciences, UK). Following sample loading, HisTrap™ columns were washed with 20mM phosphate buffer, pH 7.4 prior to elution of the sample with the same buffer containing a higher concentration of imidazole (20mM phosphate, 6M guanidinium chloride, 0.5M NaCl, 500mM imidazole, pH 7.4) during which phase five fractions (1mL) were collected and analysed by 1D SDS-PAGE.

6M GHCl independently (Figure 26). The preparation with GHCl was much more homogeneous when analysed by 1D SDS-PAGE with all of the protein eluting in a single 1mL fraction. However, the incompatibility of GHCl with SDS remained apparent, thus GHCl was removed from the protein preparation by dialysis against 1mM ammonium bicarbonate (Figure 27a) This resulted in a well resolved protein band on the 1D gel although several other faint bands were observed. In-gel digestion with trypsin and MALDI-ToF MS analysis of peptides confirmed the presence of QconCAT protein fragments in each band (Figure 27b). It is likely that a small amount of fragmentation of the intact protein had occurred during sample processing due to the denaturing conditions of SDS-PAGE, given that it is unlikely any higher order structure of the QconCAT protein exists (see discussion section 3.2.1). When analysed by 1D SDS-PAGE, fragments of QconCAT protein were present in very low amounts compared to the intact protein, such that they would not contribute a great deal to compromise quantification. To quantify this, densitometry was performed on the 1D gel image and 94% of the total protein abundance was contributed by the main QconCAT protein band, thus other fragments of QconCAT seen at this high protein loading on the gel were not significant. This unlabelled QconCAT preparation, solubilised in 6M GHCl and purified was also analysed by ESI-Q-ToF MS to recover the intact mass (Figure 28). The deconvoluted mass spectrum only contained minor peaks other than the true QconCAT protein (33036Da) and the mass offset could not be reconciled to a known pattern of adduction, thus it was concluded that this preparation was impacted less by salt contamination. To complete analysis of QconCAT labelled with [$^{13}C_6$]arg/[$^{13}C_6$]lys and purified in 6M GHCl, both labelled and unlabelled QconCAT proteins were digested in-solution with trypsin and analysed by MALDI-ToF MS. Mass spectra again contained a full complement of tryptic peptides of varying signal intensity and 'light' and 'heavy' peptides could be distinguished by their 6m/z offset as highlighted for tryptic Q-peptide T3 (Figure 29). To compare this labelling strategy with that of [$^{15}N$] in terms of natural abundance of the light isotope ([$^{12}C$]/[$^{14}N$]), the isotope profile of the same peptide was examined (Figure 30). Using [$^{13}C$] gives rise to a much less significant peak (3% of total) than [$^{15}N$] (9%) appearing at -1m/z below the monoisotopic peak, thus has less impact on quantification using this peptide. This can be overcome by taking into account the entire isotope profile for quantification but this may cause more of a problem for complex samples where peptide profiles may overlap.

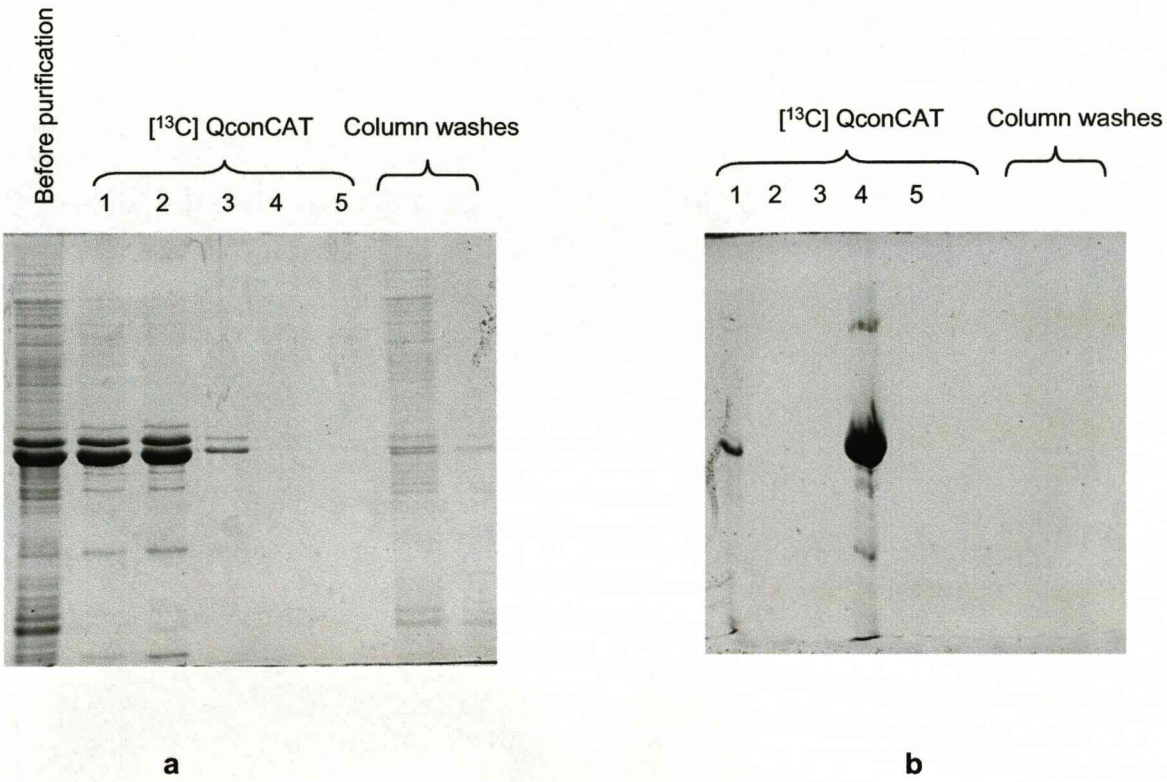



**Figure 26. Purification of [$^{13}C_6$]lys, [$^{13}C_6$]arg-labelled QconCAT protein, solubilised using 8M urea or 6M guanidinium chloride.**

[$^{13}C_6$]lys, [$^{13}C_6$]arg-QconCAT was solubilised in 20mM phosphate buffer, containing either 6M guanidinium chloride or 8M urea, 0.5M NaCl, 20mM imidazole, pH 7.4. From this solution, QconCAT proteins were separately purified by affinity chromatography using a Ni based resin (HisTrap HP Kit, Amersham Biosciences, UK). Following sample loading, HisTrap$^{TM}$ columns were washed with 20mM phosphate buffer, pH 7.4 prior to elution of the sample with the same buffer containing a higher concentration of imidazole (20mM phosphate, 6M guanidinium chloride or 8M urea, 0.5M NaCl, 500mM imidazole, pH 7.4) during which phase five fractions (1mL) were collected and analysed by 1D SDS-PAGE; a) purified in urea, b) purified in guanidinium chloride.

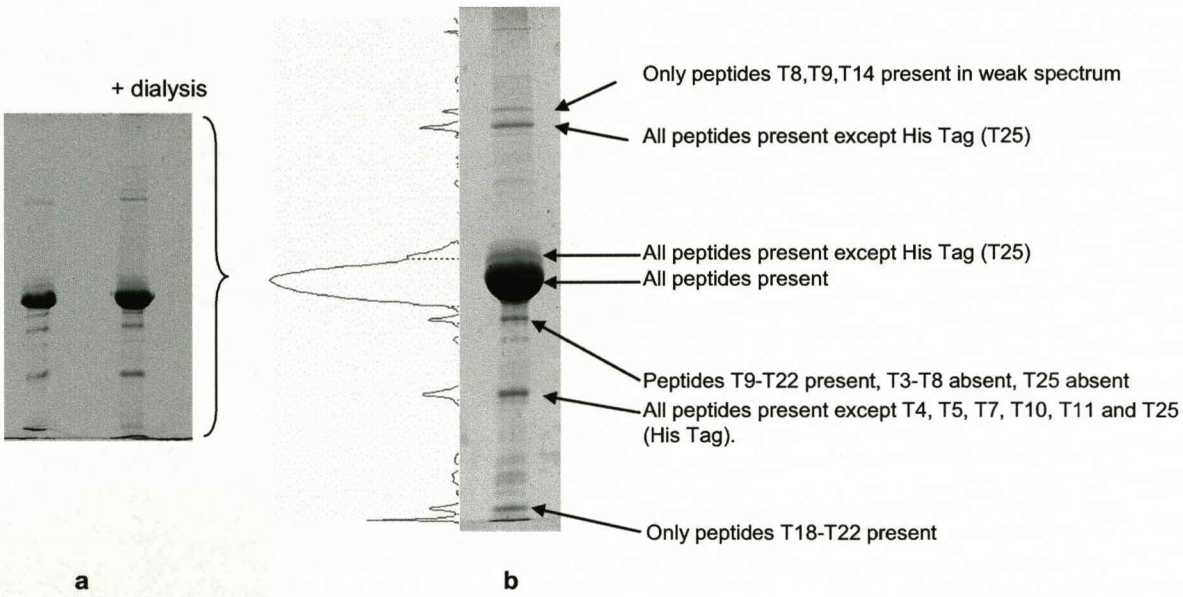



**Figure 27. Purification of [$^{13}C_6$]lys, [$^{13}C_6$]arg-QconCAT, solubilised in 6M guanidinium chloride and de-salted by dialysis.**
[$^{13}C_6$]lys, [$^{13}C_6$]arg-QconCAT was solubilised and purified in 6M guanidinium chloride. Fractions containing QconCAT protein were pooled and de-salted by three rounds of dialysis against 100 volumes 10mM ammonium bicarbonate, pH 8.5 for 2h using fresh buffer each time prior to analysis by 1D SDS-PAGE (a). The 1D SDS-PAGE gel image was analysed by densitometry to represent the protein abundance contributed by each visible band; this is presented to the left of the 1D analysis in b. For protein identification, plugs were excised from all major visible bands of QconCAT analysis after dialysis, and digested overnight in-gel with trypsin. Peptides were analysed by MALDI-ToF MS and Q-peptides identified in mass spectra were recorded (b).

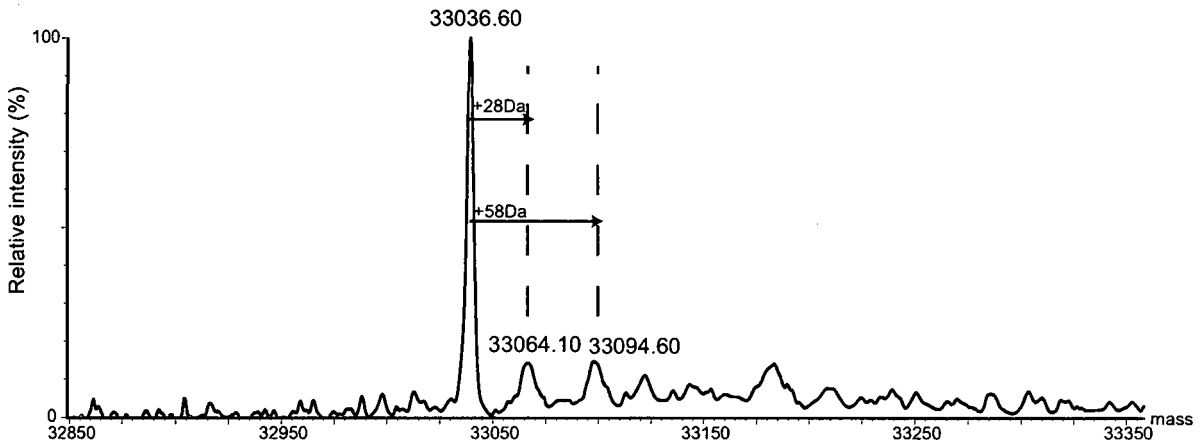



**Figure 28. Unlabelled QconCAT protein purified in 6M guanidinium chloride.**
QconCAT protein, solubilised and purified in 6M guanidinium chloride, was diluted 100 fold with 1% (v/v) formic acid and the intact mass was determined by ESI-Q-ToF MS. Mass spectra were acquired over the m/z range 700-1800 at 10% collision energy with sample cone voltage 55V and capillary voltage 1900V. The combined mass spectrum (100 scans) was deconvoluted using the MaxENT 1 maximum entropy algorithm (MassLynx) between 32800 and 33400Da at a resolution of 0.1Da/channel to recover the true mass of the protein.

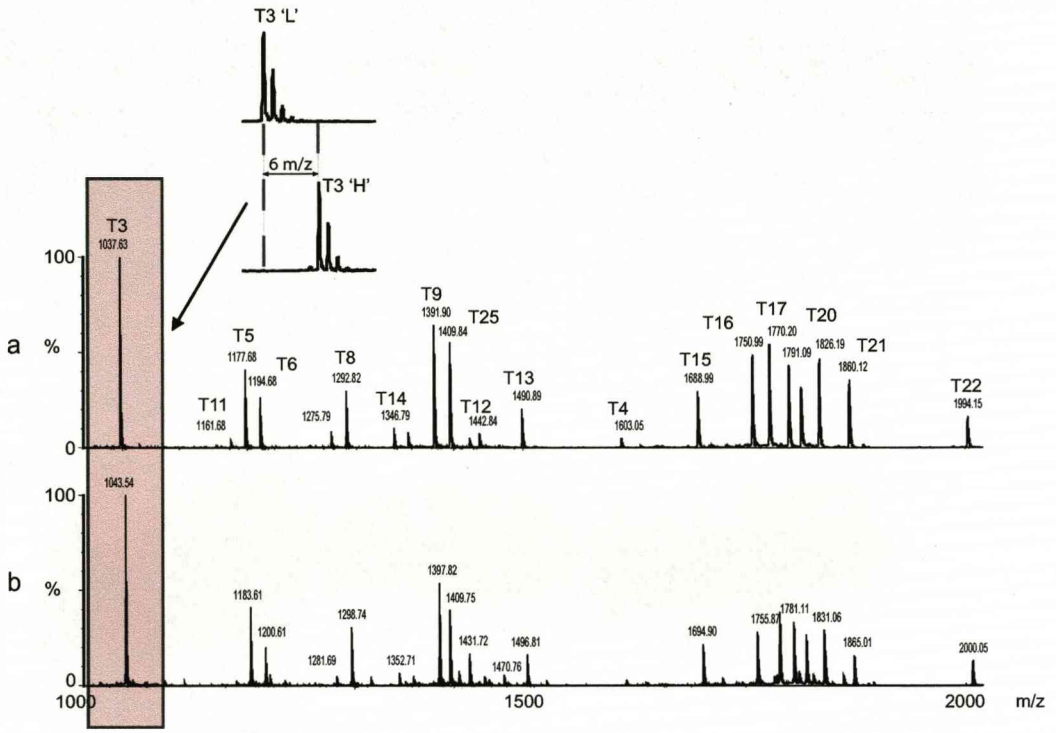



**Figure 29. Distinction between unlabelled and [$^{13}C_6$]lys, [$^{13}C_6$]arg-labelled QconCAT peptides in MALDI-ToF mass spectra.**

Unlabelled (a) and [$^{13}C_6$]lys, [$^{13}C_6$]arg-labelled (b) QconCAT proteins were diluted to 5µM in 50mM ammonium bicarbonate and digested with trypsin (20:1 substrate:protease) at 37°C for 24h. Peptides were analysed by MALDI-ToF MS. Labelled peptides are 'heavier' in mass than their unlabelled counterparts by 6Da due to the labelled C-terminal amino acid. A region of the spectrum containing the Q-peptide T3 is highlighted and this region is expanded and inserted above the main spectrum, indicating the mass offset between 'heavy' and 'light' peptides.

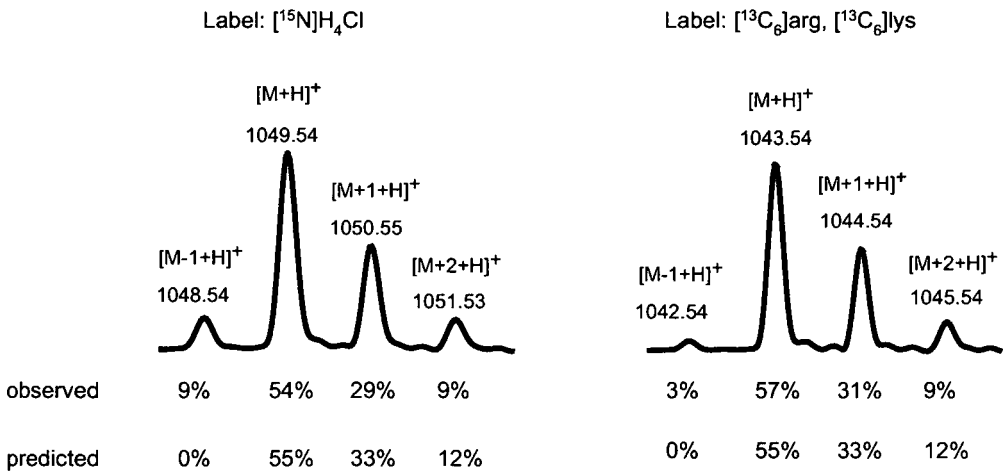



**Figure 30. Isotope distribution of a Q-peptide labelled with [$^{13}C_6$]-arg/lys and [$^{15}N$] in MALDI-ToF mass spectra.**
[$^{15}N$]-labelled and [$^{13}C_6$]lys, [$^{13}C_6$]arg-labelled QconCAT proteins were diluted to 5µM in 50mM ammonium bicarbonate and digested with trypsin (20:1 substrate:protease) at 37°C for 24h prior to analysis by MALDI-ToF MS. Zoomed regions showing the peptide envelope for a single peptide (T3; GFLIDGYPR) are illustrated, including the percentage contributed by each peak.

### 5.1.3 Proteomic analysis of chicken skeletal muscle soluble proteins

The QconCAT was designed to include surrogate peptides for quantification of twenty chicken skeletal muscle proteins (Table 3, section 1.5.2). As chicken skeletal muscle matures post-hatch, the protein distribution in the tissue changes dramatically from a large number of proteins that are expressed in similar amounts at hatch to relatively few, high abundant proteins after 30d of growth (Figure 31). From previous identification studies (Doherty *et al.,* 2004) and peptide mass fingerprinting of in-gel digestion with trypsin (Table 4; supplementary figures 1-22)), the most abundant proteins present in the soluble fraction of chicken skeletal muscle at this stage are predominantly glycolytic enzymes. Other proteins, notably actin, have disappeared from the soluble fraction of muscle by 10d of growth, presumably reflecting repartitioning and assembly of the myofibrillar apparatus. Finally, serum proteins are detectable in muscle preparations at hatch, but rapidly disappear during development. This change is most likely ascribed to the increased exclusion of interstitial fluid as the muscle develops (McLean *et al.,* 2004). The complexity of the sample is clearly observed from a MALDI-ToF mass spectrum of an entire in-solution digest of the soluble fraction of chicken skeletal muscle with trypsin (Figure 32). Only the most abundant peptides that ionise well are identified in this spectrum although the rich baseline is indicative of a complex sample with such an abundance of peptides that most are unable to be resolved. This is also reflected in MALDI-ToF mass spectra from in-gel digestion with trypsin of abundant proteins (supplementary figures 1-22), in which numerous ions cannot be assigned to the abundant protein indicating the presence of multiple proteins in the same location on the 1D gel.

## 5.2 Proteolysis of QconCAT and Analyte Proteins

### 5.2.1 Proteolysis of QconCAT

For absolute quantification of proteins using surrogate peptides as internal standards, complete proteolysis is essential. For the QconCAT method, the amount of the representative peptide selected for each protein is used to report on the amount of protein present. As such, incomplete cleavage would cause an under representation of protein amount. As the QconCAT protein is to be digested with trypsin to release the internal standard peptides, it is vital that this protein is digested efficiently and quickly. It was not expected that the three dimensional structure of the recombinant QconCAT protein would impede proteolysis as this is not a biological protein. To investigate the propensity of the QconCAT protein to be digested with the protease trypsin, QconCAT was digested in-solution with trypsin (20:1 substrate to protease)

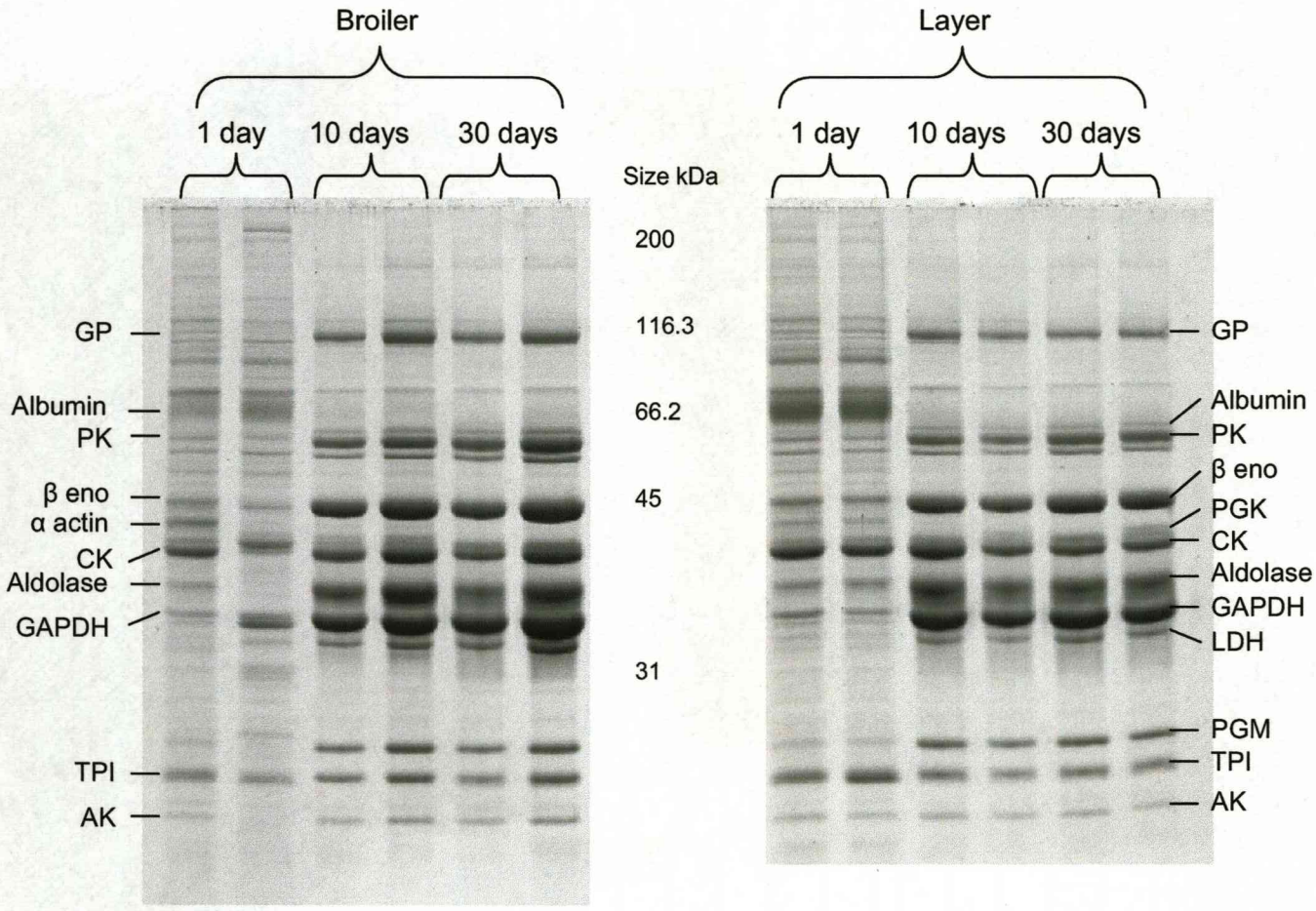



**Figure 31. Chicken skeletal muscle soluble protein expression in broiler and layer birds over 30 days of growth.**
Chickens (ISA Brown layer and Ross 308 broiler) were grown to 30d post hatch and animals were culled at 1, 3, 5, 10, 20 and 30d upon which pectoralis muscle was collected (the above procedures were performed at the Roslin Institute, Edinburgh, UK). To isolate the soluble fraction of chicken skeletal muscle, 100mg breast tissue was homogenised in 0.9mL 20mM sodium phosphate buffer, pH7.0 containing protease inhibitors (Complete Protease Inhibitors, Roche, Lewes, UK). This was centrifuged at 15,000 x *g* for 45 minutes at 4°C. The supernatant fraction, containing soluble protein, was then removed. This was repeated, homogenising the insoluble fraction in the same volume of sodium phosphate and the pooled supernatant fractions were used for all analyses. 10µg soluble protein samples (volume 5-10µL) from birds of different strains and ages were analysed by SDS-PAGE. For protein identification, gel plugs were excised and digested with trypsin and peptides were analysed by MALDI-ToF MS. For details, see table 4 (overleaf) and individual mass spectra in supplementary figures (1-22).

| | Protein | MOWSE score | Swissprot accession # | Molecular weight (Da) | First species | # peptides matched | % coverage |
|---|---|---|---|---|---|---|---|
| 1d | GP | 76 | Q7ZZK3 | 98566 | *Gallus gallus* | 11 | 15 |
| | Albumin | 113 | P19121 | 71800 | *Gallus gallus* | 12 | 23 |
| | Pyruvate kinase | 113 | P00548 | 57847 | *Gallus gallus* | 19 | 33 |
| | B eno | 134 | P07322 | 46839 | *Gallus gallus* | 10 | 33 |
| | A actin | 72 | P68139 | 42051 | *Gallus gallus* | 6 | 24 |
| | Creatine kinase | 148 | P00565 | 43301 | *Gallus gallus* | 12 | 31 |
| | Aldolase | | | | | | |
| | GAPDH | 127 | P00356 | 35681 | *Gallus gallus* | 9 | 40 |
| | TPI | 186 | P00940 | 26620 | *Gallus gallus* | 13 | 43 |
| | AK | 101 | P05081 | 21683 | *Gallus gallus* | 7 | 41 |
| | | | | | | | |
| 30d | GP | 105 | Q7ZZK3 | 98566 | *Gallus gallus* | 12 | 15 |
| | Albumin | 51 | P19121 | 71800 | *Gallus gallus* | 3 | 7 |
| | PK | 195 | P00548 | 57978 | *Gallus gallus* | 17 | 38 |
| | B eno | 75 | P07322 | 46839 | *Gallus gallus* | 10 | 21 |
| | PGK | 88 | P51903 | 44557 | *Gallus gallus* | 7 | 24 |
| | CK | 187 | P00565 | 43301 | *Gallus gallus* | 19 | 45 |
| | Aldolase | | | | | | |
| | GAPDH | 135 | P00356 | 35681 | *Gallus gallus* | 16 | 52 |
| | LDH A | 189 | P00340 | 36491 | *Gallus gallus* | 22 | 46 |
| | PGM | 109 | Q5ZLN1 | 28749 | *Gallus gallus* | 10 | 40 |
| | TPI | 176 | P00940 | 26620 | *Gallus gallus* | 13 | 48 |
| | AK | 101 | P05081 | 21683 | *Gallus gallus* | 7 | 41 |

**Table 4. Identification of chicken skeletal muscle soluble proteins by peptide mass fingerprinting.**

Chicken skeletal muscle soluble proteins were analysed by 1D SDS-PAGE and digested in gel with trypsin. Peptides were analysed by MALDI-ToF MS and monoisotopic masses were entered into the MASCOT search engine. Data were searched against the database MSDB for taxonomy: Chordata, variable modifications: oxidation of methionine, protease: trypsin, missed cleavages: 1, peptide tolerance: 250ppm. MOWSE scores above 65 at probability level, p=0.05 were accepted as confident matches. Mass spectra corresponding to identified proteins are presented in supplementary figures (1-22). NB. Muscle type aldolase (A) in chicken is not sequenced and incorporated into the database Swissprot. The Q-peptide incorporated for this protein was taken from liver type aldolase (B) and thus is not represented in the soluble fraction of chicken skeletal muscle. The identity of this protein was discussed and confirmed by Dr. J. Hayter (Hayter *et al.*, 2003).

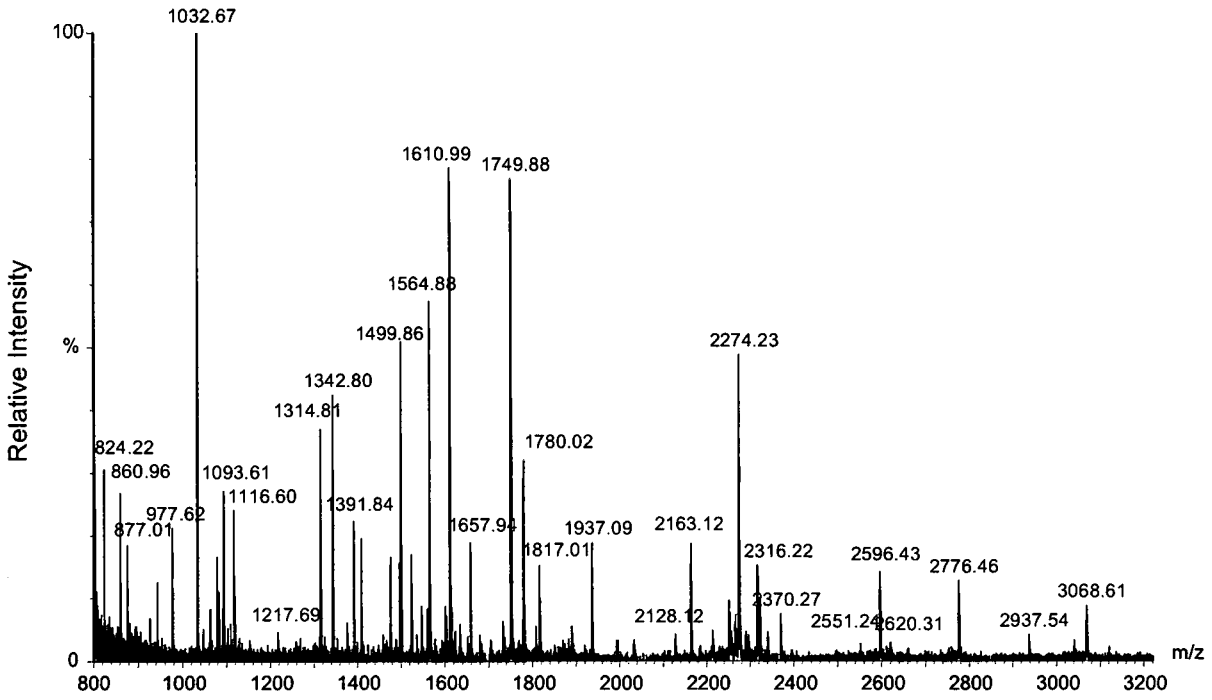
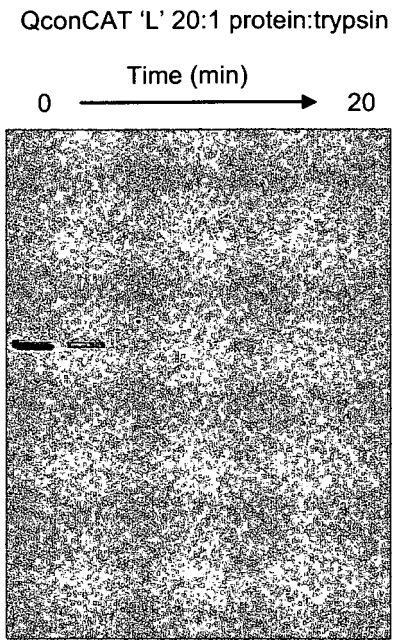



**Figure 32. Chicken skeletal muscle soluble proteins digested in-solution with trypsin.**

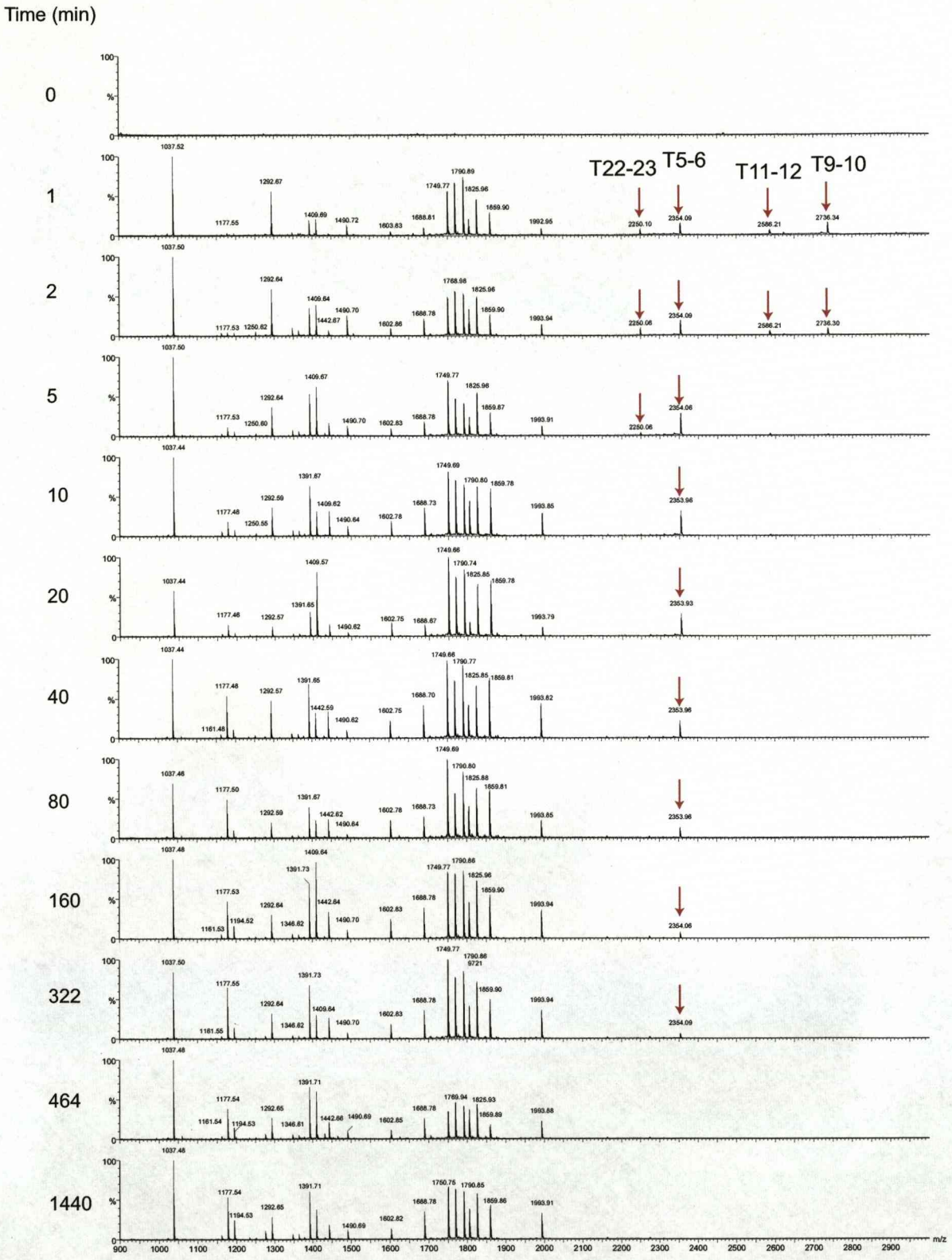
Chicken skeletal muscle soluble proteins from a 30d broiler were diluted 10 fold with 50mM ammonium bicarbonate, prior to addition of trypsin (20:1 substrate:protease). The reaction mixture was incubated at 37°C for 24h after which the digest was incubated with additional trypsin (20:1 substrate:protease) to achieve complete digestion and 1µL was analysed by MALDI-ToF MS.

and the reaction was stopped at selected time points by the addition of digesting material to 10% (v/v) formic acid. The disappearance of the intact protein was monitored by 1D SDS-PAGE by drying down the digested mixture to remove the acid and reconstituting in reducing sample buffer immediately prior to loading. The QconCAT protein had completely disappeared from the gel after two minutes of digestion with no intermediate fragments visible after this time (Figure 33a). The appearance of limit peptides was also monitored by MALDI-ToF MS, with all peptides present in the spectrum after one minute of digestion with trypsin (Figure 33b). The disappearance of incomplete digestion products greater than 2000Da can also be observed as digestion reaches completion around 8h (464 min). This rapid proteolysis is to be expected as the QconCAT protein is not a biological entity, is not expected to fold into a complex 3D structure and therefore does not contain regions inaccessible to trypsin cleavage (Hubbard *et al.,* 1991). When the trypsin was reduced to much lower levels (100:1 substrate to protease) and the digestion reaction was sampled at very short time intervals, there was some evidence for the appearance of partially digested intermediates, although in-gel digestion of these bands (1-7) and subsequent MALDI-ToF MS analysis of peptides demonstrated that each band comprised multiple species (Figure 34), consistent with simultaneous tryptic attack on all scissile bonds at very similar rates. This was confirmed by digesting QconCAT protein in-solution with trypsin at the same enzyme to protein ratio (1:100), taking early time points where digestion was stopped by the addition of 10% (v/v) formic acid and analysing the resulting digested mixtures on a 30cm 1D gel, rather than the smaller 7cm gel. When separated through a larger gel, there were an increased number of visible bands which were presumed to have overlapped on the smaller gel (Figure 35). To investigate the route of proteolysis further, peptides from a low concentration trypsin digestion of QconCAT protein in-solution were analysed by MALDI-ToF MS over 24h. This revealed that some cleavage sites were favoured, although all expected peptides are present after 30m of digestion under these conditions (Figure 36). Additionally, some cleavage sites remain resistant to trypsin digestion throughout the time course of this experiment, for example T5-6 ($[M+H]^+$ 1253.96m/z; NLAPYSDEL**R**GDQLFTATEGR). It is predicted that glycine may have an association with missed cleavage in tryptic peptides, although not specifically C-terminal to the arg/lys residue (Siepen *et al.,* 2007), thus there is no obvious reason for this –ArgGly- cleavage to be problematic. In the context of digestion of the QconCAT protein for absolute quantification of analyte proteins, proteolysis was so fast even at low concentrations of protease, that further investigation of the mechanism for digestion in this case was unnecessary.




a

**Figure 33. Proteolysis of QconCAT with trypsin.**
For QconCAT digestion, 150µg protein was digested with trypsin at a ratio of trypsin to protein of 1:20 and stopped at selected time points after addition of enzyme by removing 15µL (containing 3µg protein) and adding to an equal volume of 10% (v/v) formic acid. The fractions were subsequently stored at -20°C until the end of the time course. 25µL of each fraction were dried down in a vacuum centrifuge and reconstituted in 10µL reducing sample buffer prior to analysis by 1D SDS-PAGE (a). 1µL of each fraction was analysed by MALDI-ToF MS (b; overleaf).


Time (min)
0
1
2
5
10
20
40
80
160
322
464
1440
T22-23
T5-6
T11-12
T9-10
m/z

b

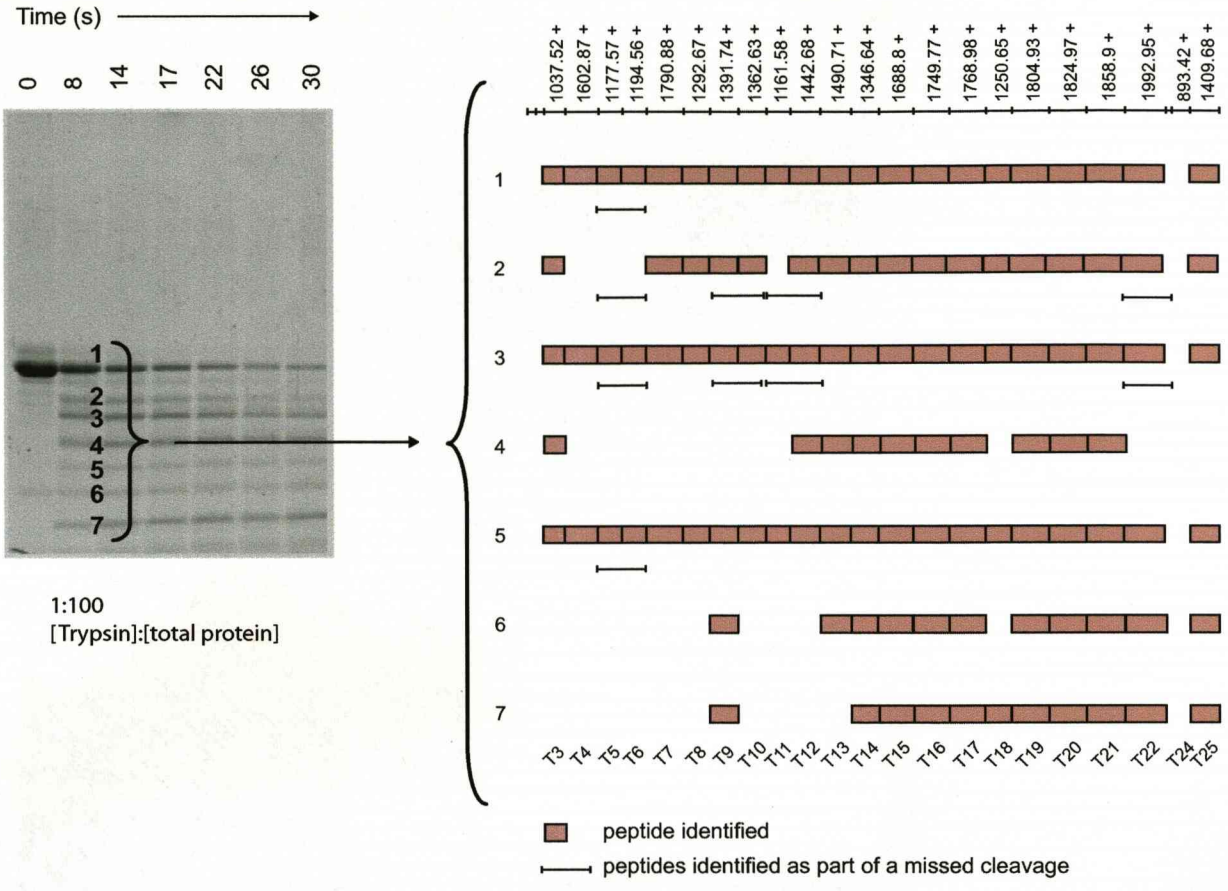



**Figure 34. Proteolysis of QconCAT and diagnostic peptide mass fingerprinting of peptides.**
QconCAT protein (150µg) was digested with trypsin at an enzyme:protein ratio of 1:00. The digestion was stopped at selected time points after addition of enzyme with 10% (v/v) formic acid. For gel electrophoresis, fractions from QconCAT protein digestion were dried down in a vacuum centrifuge and reconstituted in 10µL reducing sample buffer prior to analysis. Protein bands observed after 8s of proteolysis were digested in-gel with trypsin and peptides were analysed by MALDI-ToF MS. Peptides observed in each MALDI-ToF mass spectrum are indicated by a coloured block, peptides identified joined by a missed cleavage site are indicated with a black line under each peptide map.

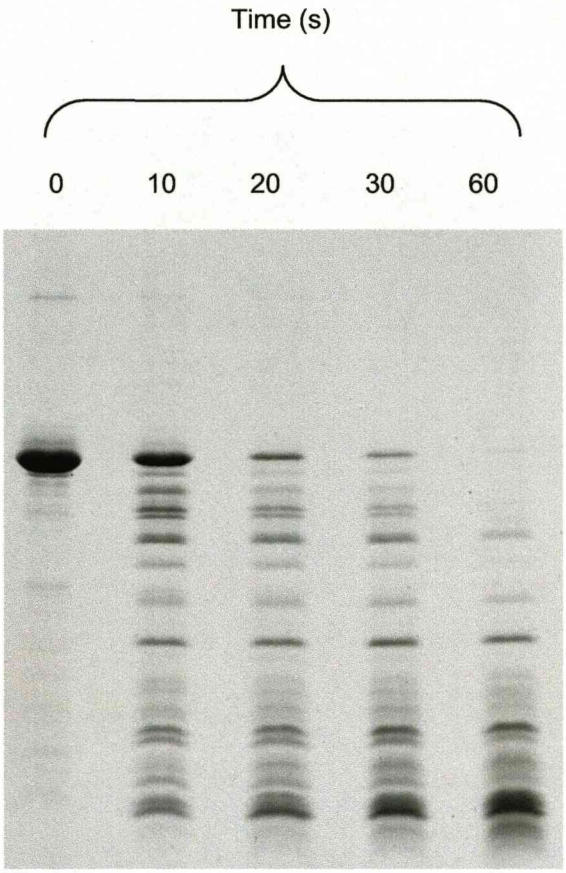



**Figure 35. Proteolysis of QconCAT and extensive separation of protein fragments by SDS-PAGE.** QconCAT protein (150µg) was digested with trypsin at an enzyme:protein ratio of 1:00. The digestion was stopped at selected time points after addition of enzyme with 10% (v/v) formic acid. For gel electrophoresis, fractions from QconCAT protein digestion were dried down in a vacuum centrifuge and reconstituted in 10µL reducing sample buffer prior to loading onto a 30cm 1D gel.

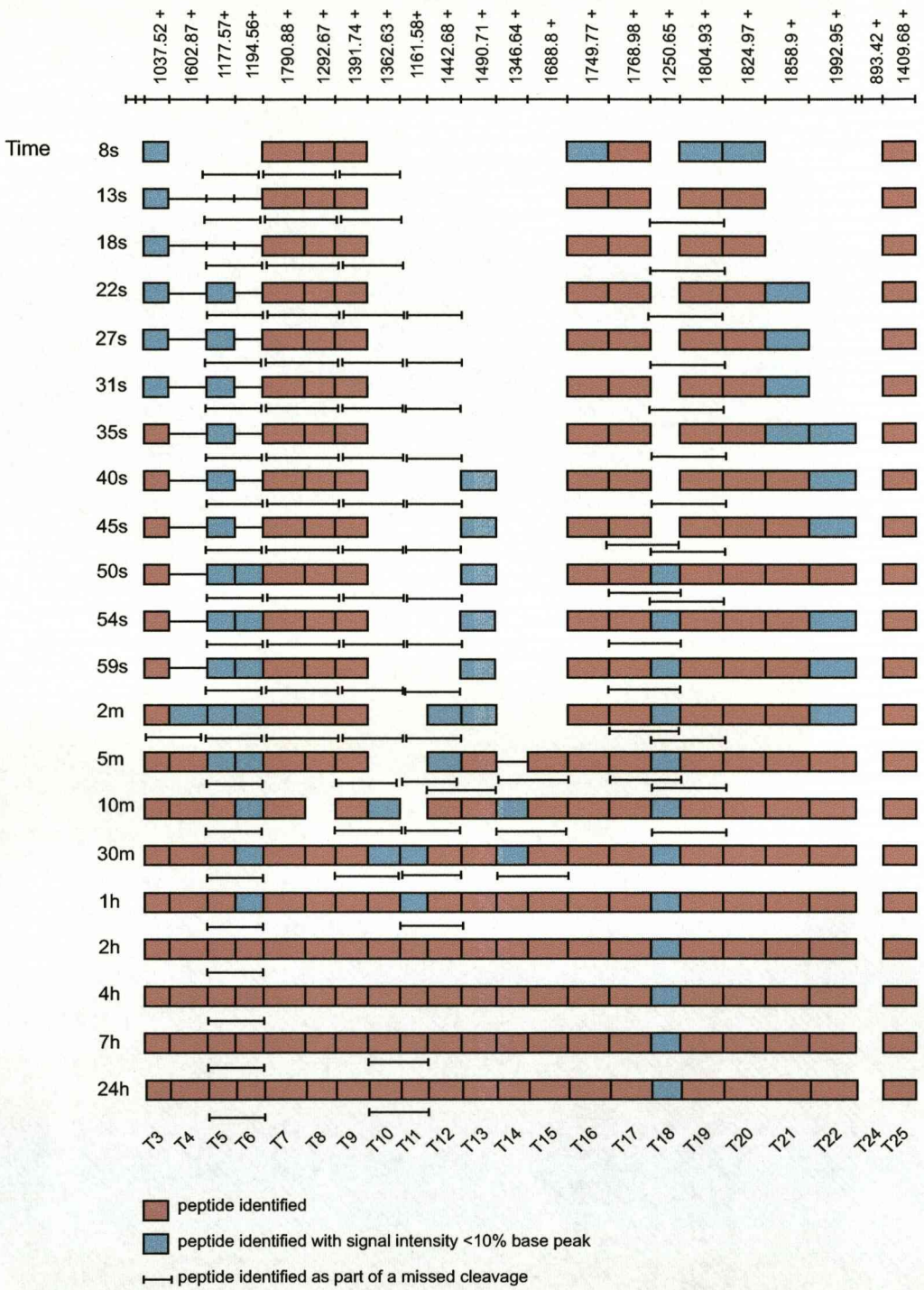



**Figure 36. Proteolysis of QconCAT and analysis of peptide release.** .
QconCAT protein (150µg) was digested with trypsin at an enzyme:protein ratio of 1:00. The digestion was stopped at selected time points after addition of enzyme with 10% (v/v) formic acid. 1µL digested material was analysed by MALDI-ToF MS and peptides present in each spectrum are indicated in pink (high abundance ions), blue (low abundance ions; signal intensity less than 10% of the base peak) and those present as part of a missed cleavage are indicated by a black line underneath each peptide map.

### 5.2.2 Proteolysis of chicken skeletal muscle soluble proteins

In contrast to the QconCAT protein, the analyte proteins were expected to have much more protracted proteolysis reactions due to the effects of higher order structure and the context of complex mixtures of proteins limiting availability of trypsin cleavage sites. For the soluble proteins of chicken skeletal muscle, in-solution digestion with trypsin at a ratio of 20:1 (substrate:protease), with the reaction stopped at selected time points by addition of 10% (v/v) formic acid, indicated that many proteins were digested slowly, and even after 24h, undigested proteins were clearly visible by 1D SDS-PAGE including beta enolase (β eno), creatine kinase (CK) and triose phosphate isomerase (TPI) (Figure 37a). If a low concentration (10% v/v) of acetonitrile was included in the digestion reaction, proteolysis was much faster and all bands had disappeared by 24h (Figure 37b). The reasons for this are unclear but it is likely that the actions of organic solvents during proteolysis are to facilitate denaturation, particularly of secondary structure, and increase solubility of native proteins exposing the active site (Russell *et al.,* 2001). When the analyte protein mixture was denatured by heating to 60°C for 1h before digestion, the disappearance of intact protein bands when the products were analysed by 1D SDS-PAGE suggested that the loss of higher order structure of the substrate proteins caused the digestion reaction to be essentially complete within 30min (Figure 38c). This was directly compared to analyte protein digestion at an enzyme to substrate ratio of 1:20 and with 10% (v/v) added acetonitrile for the same biological sample (Figure 38a&b). Again, addition of organic solvent seemed to accelerate the reaction (although this did not seem to be complete after 24h) and the majority of proteins had disappeared from the gel following denaturation by heating prior to 30min proteolysis with trypsin. Although protein denaturation by heating permits much more efficient digestion with trypsin, this could also cause proteins to precipitate out of solution and indeed the initial protein profile prior to addition of enzyme is diminished in the heat treated sample even though no precipitation was observed (Figure 38c).

For absolute protein quantification using stable isotope labelled internal standard peptides, complete proteolysis is vital as the peptide used for quantification must be completely released from the protein of interest in order to ascertain its absolute amount. As a tool to quantify this release, the rapidly and efficiently digested [$^{13}C_6$]arg/[$^{13}C_6$]lys-labelled QconCAT protein was digested to completion in-solution with trypsin. As each peptide exists in equal amounts within the QconCAT protein, a known amount of digested material may also be used as an internal standard to quantify the release of each peptide from unlabelled QconCAT protein when digested with trypsin. To achieve this, a known amount of 'light' QconCAT digesting material

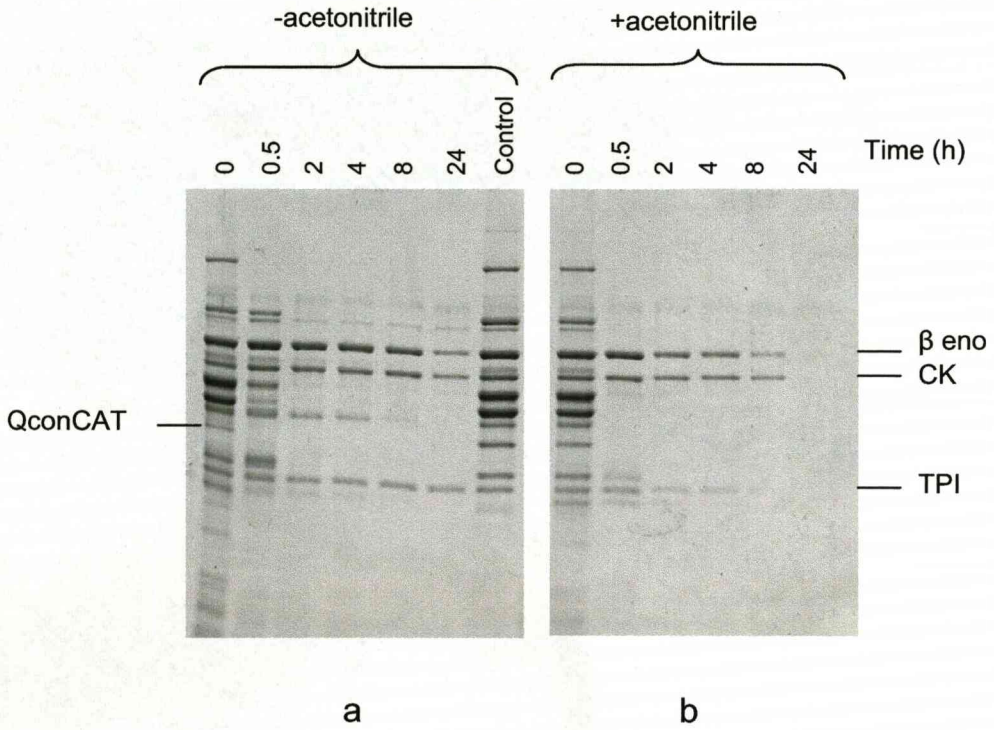



**Figure 37. In-solution proteolysis of chicken skeletal muscle soluble proteins with trypsin and the effect of added acetonitrile.**

50µg soluble protein from skeletal muscle of a 30d layer chicken was added to 5µg labelled QconCAT protein and digested with trypsin at a ratio of trypsin to total protein of 1:20 and stopped at selected time points after addition of enzyme by removing 25µL (containing 6µg protein) and adding to an equal volume of 10% (v/v) formic acid. The fractions were subsequently stored at -20°C until the end of the time course. For gel electrophoresis, fractions were dried down in a vacuum centrifuge and reconstituted in 10µL reducing sample buffer prior to analysis by 1D SDS-PAGE (a). Analyte proteins were also digested under the same conditions in a solution containing 10% (v/v) acetonitrile (b).

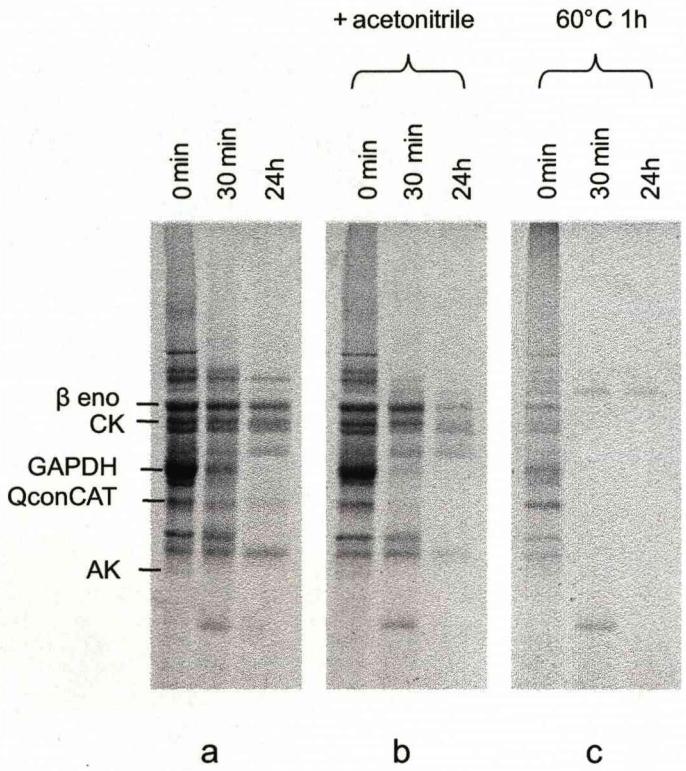



**Figure 38. In-solution proteolysis of chicken skeletal muscle soluble proteins with trypsin, the effect of added acetonitrile, and protein denaturation prior to the addition of protease.** 50µg soluble protein from skeletal muscle of a 30d broiler chicken was added to 5µg labelled QconCAT and digested with trypsin at a ratio of trypsin to total protein of 1:20 and stopped at 0min, 30min and 24h after addition of enzyme by removing 25µL (containing 6µg protein) and adding to an equal volume of 10% (v/v) formic acid. The fractions were subsequently stored at -20°C until the end of the time course. For gel electrophoresis, fractions were dried down in a vacuum centrifuge and reconstituted in 10µL reducing sample buffer prior to analysis by 1D SDS-PAGE (a). Analyte proteins were also digested under the same conditions in a solution containing 10% (v/v) acetonitrile (b) and with addition of enzyme following a 1h incubation of the protein at 60°C (c).

was removed at selected time points during proteolysis with trypsin at 37°C and added to an equal volume of 10% (v/v) formic acid containing an equal amount of pre-digested QconCAT 'heavy' peptides. Peptides at each time point were analysed by MALDI-ToF MS and the extent of digestion was calculated by comparing signal intensity of 'heavy' (completely digested) to 'light' (digesting) peptides during proteolysis (Figure 39). This confirmed and quantified the rapid digestion of the QconCAT protein, for all peptides. To apply this technique for chicken skeletal muscle soluble proteins, extended digestion reactions with preparations from 1d and 30d birds were devised. As reported previously, the protein expression profiles of these two preparations are dramatically different (Figure 31), providing alternative environments for proteolysis. The protein preparations were digested with trypsin in-solution, without treatment or after denaturation at 60°C for 1h, and the reaction was stopped at selected time points by adding a known proportion to 10% (v/v) formic acid containing a known amount of pre-digested QconCAT 'heavy' peptides. The appearance of analyte peptides used for quantification was monitored by MALDI-ToF MS and quantified (nmol/g tissue) by QconCAT; the amount of proteolysis was quantified as nmol/g tissue to put this experiment into the context of absolute quantification and to verify that the final value reached for each protein is consistent with that achieved for the biological study. To analyse the initial stages of proteolysis, data for proteolysis of native and denatured proteins are presented up to 500min (Figure 40), in addition to extended digestion times (24h+; Figure 41) to emphasize that complete digestion is achieved, and the same quantification value is reached, irrespective of the initial state of the analyte protein preparation. In all instances, the analyte proteins were digested between 1.3 (AK) and 86 (β eno) times faster after denaturation, and in some instances (for example, GAPDH from one day muscle) the rate of digestion was very similar. This is consistent with a model for proteolysis of the native protein in which the initial proteolytic attack exerts a de-stabilising effect on the remaining structure, such that the rate of proteolysis is increased; the initial proteolysis is effectively rate limiting. The context in which proteins are digested also seems to play an important role; in the highly specialised 30d muscle sample, there was virtually no digestion of creatine kinase, until 20 hours of proteolysis. Indeed, for all proteins studied, the rate of proteolysis of native proteins was diminished in the 30d muscle sample, suggesting that the acute specialisation of this tissue, leading to a predominance of relatively few proteins, might introduce other factors that impede digestion, such as aggregation into supramolecular assemblies or partial inhibition of the trypsin.

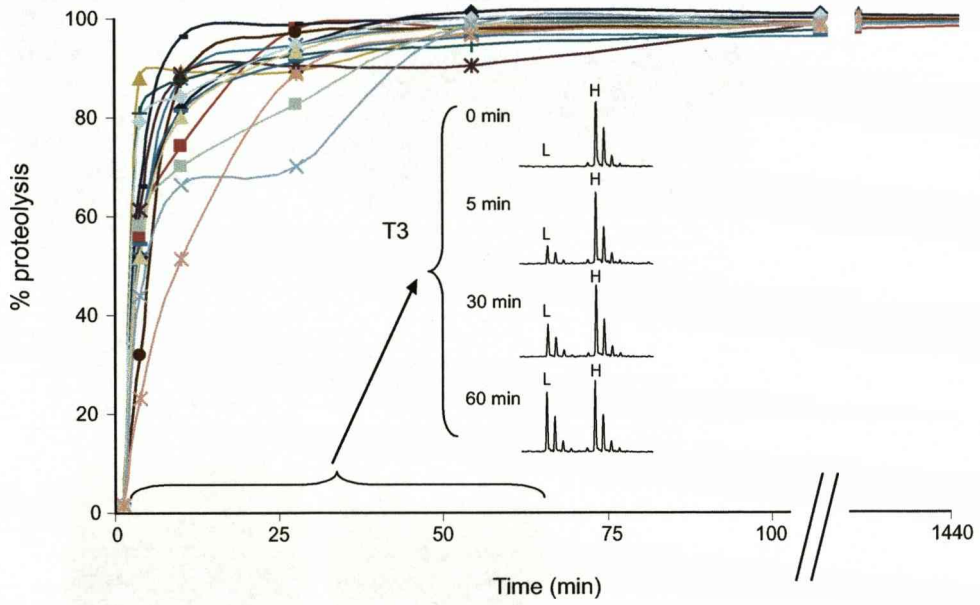



**Figure 39. Absolute quantification of unlabelled QconCAT proteolysis.**
5µg unlabelled QconCAT protein were digested in-solution with trypsin at a ratio of trypsin to protein of 1:20 and the reaction was stopped at selected time points during 24h incubation at 37°C by removing 0.5µg protein and adding an equal volume of 10% (v/v) formic acid containing 0.5µg pre-digested QconCAT peptides. Each fraction was analysed by MALDI-ToF MS and the relative signal intensity of unlabelled and labelled peptides was used to calculate the percentage of proteolysis. Digestion was complete (100%) when signal intensity of unlabelled and labelled peptides was equal. This was plotted for individual tryptic Q-peptides and percentage proteolysis is illustrated for the first 500min of incubation with enzyme. Inserted are zoomed mass spectra for the Q-peptide T3 illustrating the increase in abundance of unlabelled peptide during the first 60min of proteolysis.

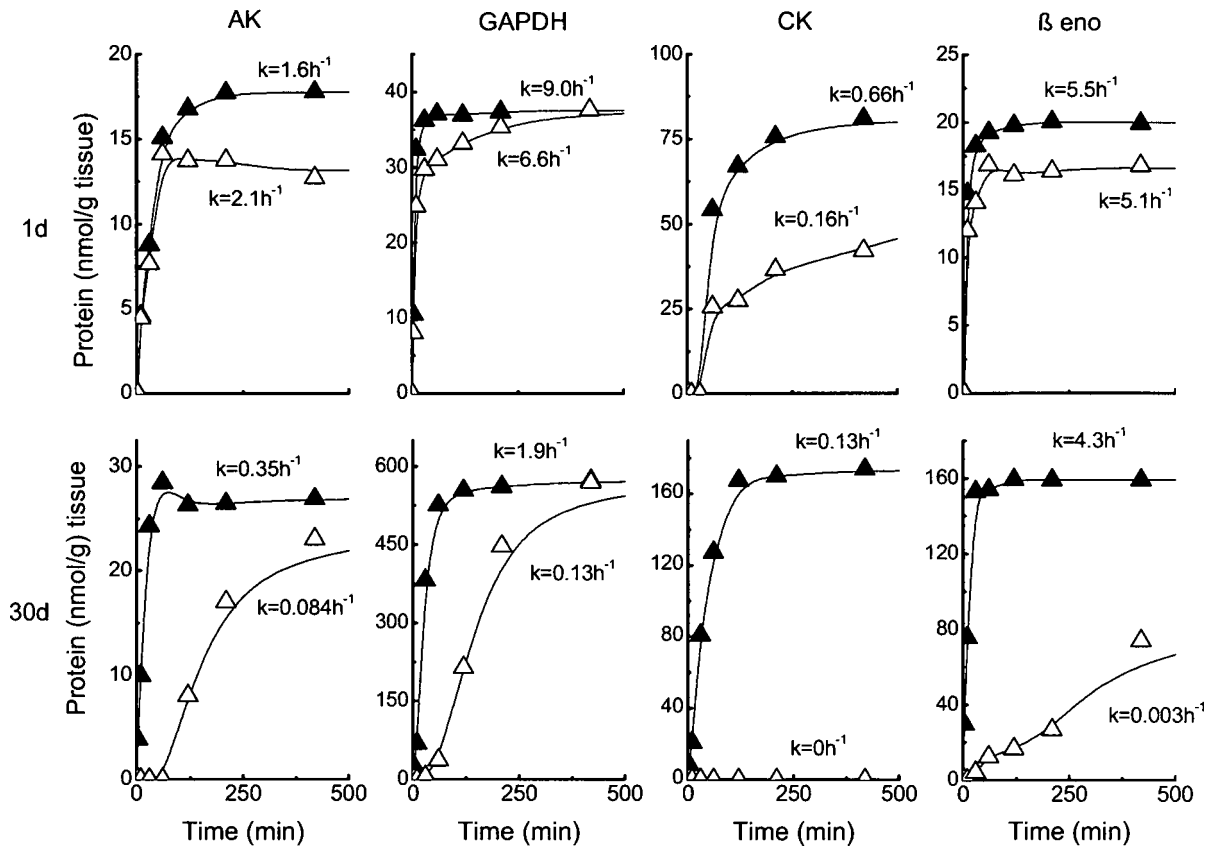



**Figure 40. Quantification of 500min proteolysis of analyte proteins with trypsin using QconCAT.** Chicken skeletal muscle soluble protein (50µg) was digested with trypsin at an enzyme:protein ratio of 1:20 and stopped at selected time points with 10% (v/v) formic acid and mixed with 0.5µg pre-digested QconCAT peptides for quantification. Each fraction was analysed by MALDI-ToF MS. This experiment was repeated using protein denatured by incubating at 60°C for 1h prior to trypsin addition for comparison. Data are presented for four individual proteins at both 1d and 30d after hatch digested over 500min with trypsin and for each, the rate constant (k) for digestion is expressed as $h^{-1}$ (from first order decay of loss of substrate, S from 100%; $S=100e^{-kt}$).

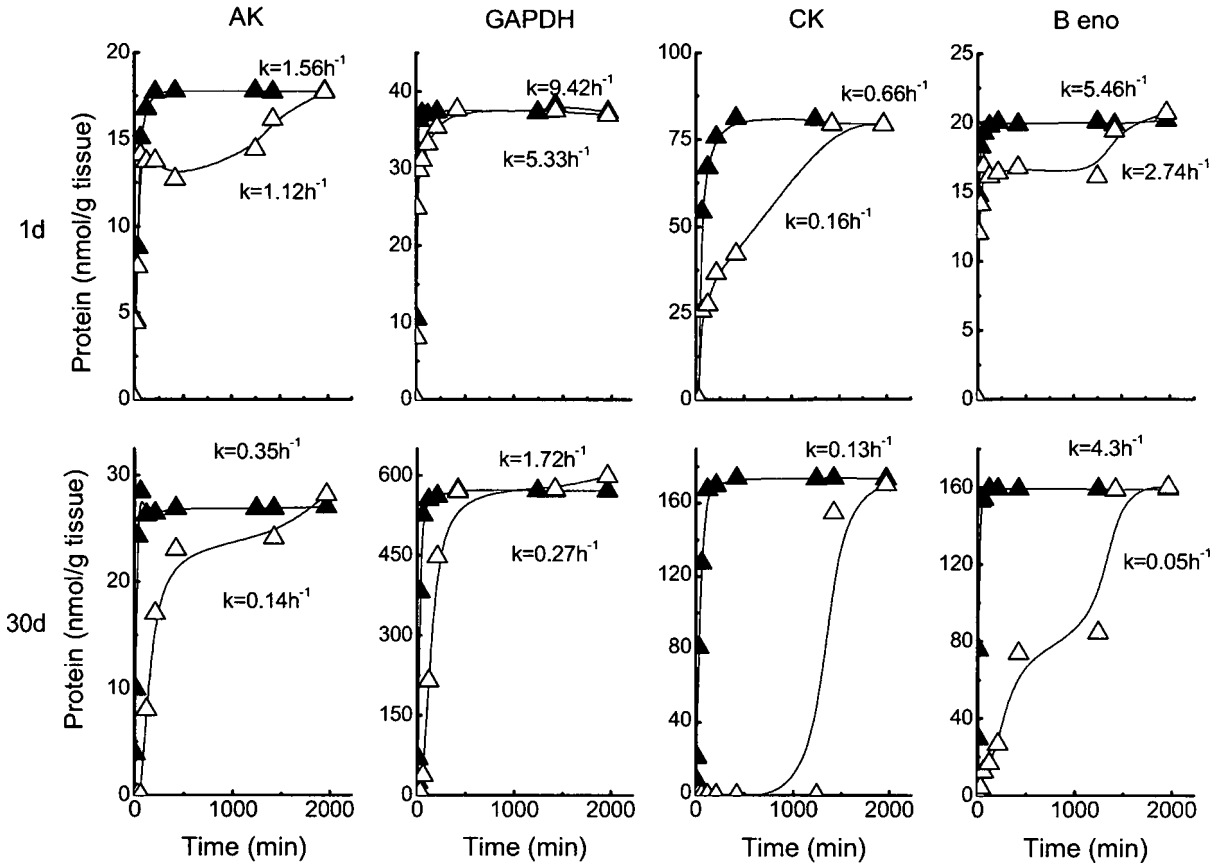



**Figure 41. Quantification of 30h proteolysis of analyte proteins with trypsin using QconCAT.**
Chicken skeletal muscle soluble protein (50µg) was digested with trypsin at an enzyme:protein ratio of 1:20 and stopped at selected time points with 10% (v/v) formic acid and mixed with 0.5µg pre-digested QconCAT peptides for quantification. Each fraction was analysed by MALDI-ToF MS. This experiment was repeated using protein denatured by incubating at 60°C for 1h prior to trypsin addition for comparison. Data are presented for four individual proteins at both 1d and 30d after hatch digested over 30h with trypsin and for each, the rate constant (k) for digestion is expressed as $h^{-1}$ (from first order decay of loss of substrate, S from 100%; $S=100e^{-kt}$)

## 5.3 Sample Complexity and Dynamic Range

### 5.3.1 Mass Spectrometry for absolute quantification using the QconCAT method

Theoretically, proteolysis of a complex proteome (for example 10,000 proteins) could generate $10^5$-$10^6$ peptides (at approximately 50 tryptic peptides per protein), the dynamic range of which will be such that only the most abundant peptides or those that ionise particularly well will be detectable. For analysis of the same sample, the most efficiently ionised peptides will vary with the use of different instruments. For reliable absolute quantification of proteins using mass spectrometry, the QconCAT method must be robust across these platforms. QconCAT was added to soluble proteins in chicken skeletal muscle from four chickens at each of six times points during growth of both broiler and layer strains in a known amount and peptides derived from in-solution digestion with trypsin were analysed using both MALDI-ToF MS and LC-ESI-Q-ToF MS for absolute quantification. Due to variation in ionisation, several peptides were only identified using one instrument, for example triose phosphate isomerase was only observed in ESI and glycogen phosphorylase in MALDI (absolute quantification for proteins achieved using different instruments is discussed in the context of the biological experiment, section 3.5). Several proteins were analysed and quantified in both instruments, permitting a direct comparison (Figure 42). Strong, linear correlation of slope 1.05 and $R^2$ 0.98 confirmed that quantification was reproducible across both platforms. This highlighted the usefulness of using a combination of instruments to gain as much information as possible from an analytical sample, although in this case identification of six proteins in ESI that had not previously been identified in MALDI was predominantly due to the reversed phase separation; discussed in section 3.3.2. It was also necessary at this stage to verify the assumption that ionisation of different locations on the target plate for MALDI-ToF MS containing the same sample will achieve the same mass spectrum following detection. To confirm this, QconCAT peptides, both unlabelled and labelled with [$^{13}C_6$]arg/[$^{13}C_6$]lys were mixed in an approximate 1:1 ratio before 1µL was spotted onto a MALDI target with an equal volume of matrix (α-cyano hydroxycinnamic acid). 15 different laser positions on a single sample well were ionised independently and 20 spectra were acquired and combined for each location. For six QconCAT peptides, the ratio of analyte ('light') to standard ('heavy') was measured and is presented as the mean±standard error of the mean (Figure 43). The observation that L:H ratios for the six peptides are not constant is most likely due to inaccuracies with mixing peptides after independent proteolysis, rather than co-digestion of proteins where sample processing conditions can be more precisely controlled. The reproducibility of MALDI ionisation across the target well is clear, thus quantification is not affected by the specific location irradiated by the

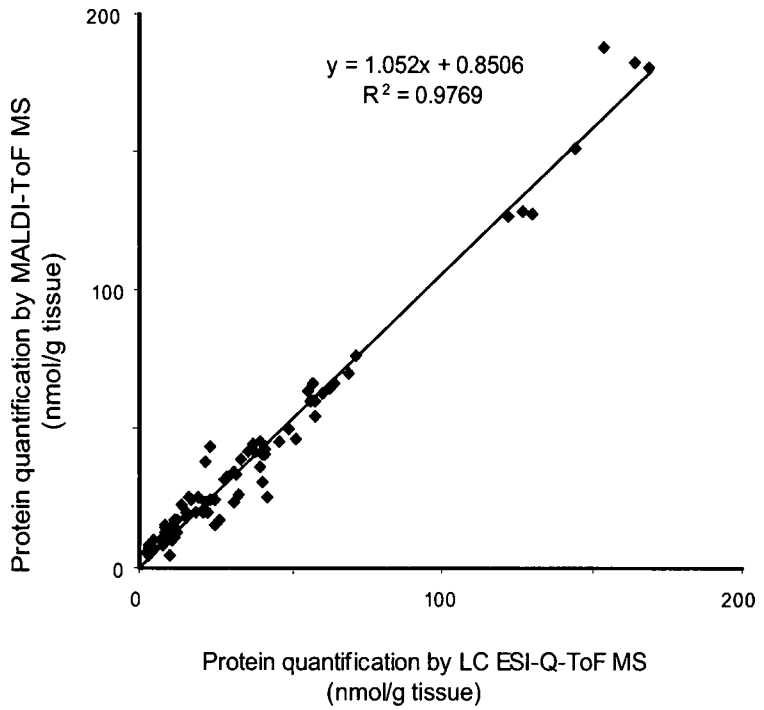



**Figure 42. Absolute quantification of chicken skeletal muscle soluble proteins by MALDI-ToF MS and LC-ESI-Q-ToF MS.**

Soluble proteins from chicken skeletal muscle (70µg, n=4, covering 1d to 30d post hatch) were individually mixed with QconCAT protein (7µg) and digested to completion with trypsin. The entire peptide mixture was analysed by MALDI-ToF MS or by nanoflow reversed phase HPLC prior to ESI-Q-ToF MS and the absolute tissue content of each of four proteins (triose phosphate isomerase, glyceraldehyde 3-phosphate dehydrogenase, beta enolase and alpha actin) was assessed from relative intensities of light (analyte) and heavy (standard) pairs. The absolute amount of each protein was compared using the alternative forms of mass spectrometric analysis.

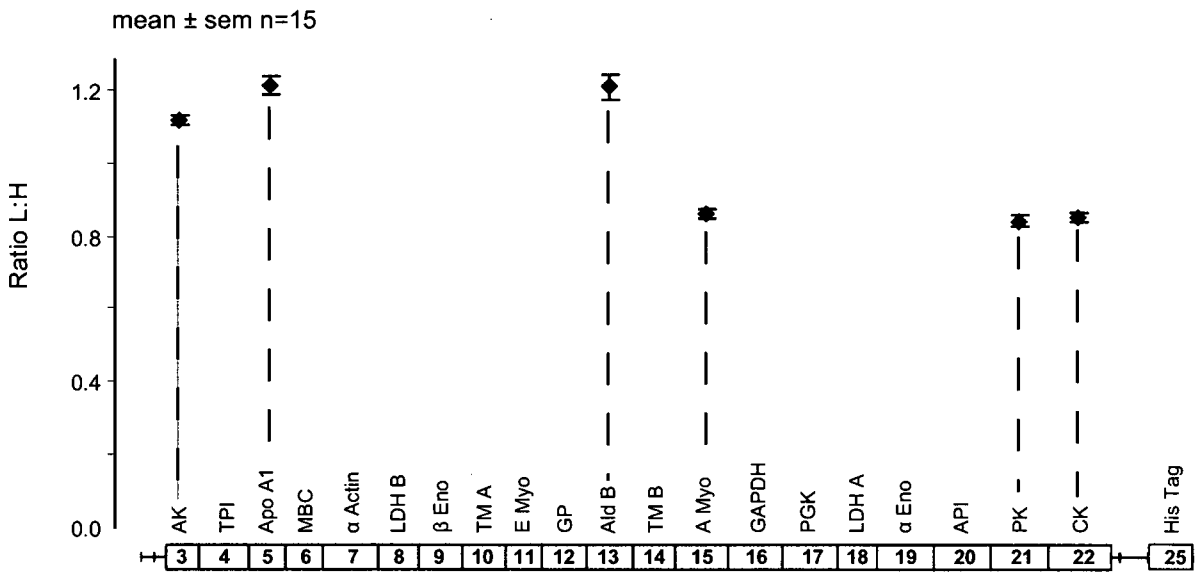



**Figure 43. Relative intensity of QconCAT labelled and unlabelled peptides acquired from 15 different locations on a MALDI target.**

QconCAT unlabelled and labelled peptides were mixed in an approximate 1:1 ratio and 1µL was spotted onto a MALDI target and mixed with 1µL matrix (α-cyano hydroxycinnamic acid). 20 MALDI spectra were acquired and combined from each of 15 randomly selected locations on the sample well and the ratio of unlabelled:labelled signal intensity was calculated, with the mean ± standard error (sem) plotted for 6 Q-peptides.

laser. Indeed for most MALDI-ToF MS acquired data, spectra are summed across various points on the target in a random format to obtain the most stable signal according to user discretion.

### 5.3.2 Challenges for data acquisition and analysis for quantification

Identifying and quantifying as many peptides as possible from a single MALDI-ToF mass spectrum is a challenge for proteomics due to the high level of sample complexity often encountered. For many peptides, a signal is observed in the mass spectrum but cannot be used for absolute quantification as it overlaps with another analyte peptide also contributing signal to the peptide envelope. The soluble fraction of chicken skeletal muscle is highly complex and contains thousands of proteins. For absolute quantification, these proteins are mixed with the QconCAT protein and co-digested with trypsin, resulting in an even more complex sample containing far too many peptides to be discretely analysed using MALDI-ToF MS. For quantification using the QconCAT method, this may be overcome in part by selecting, where possible, peptides that are known to ionise well within the analytical system of choice. This was possible for several peptides incorporated into the QconCAT as internal standards for chicken skeletal muscle soluble proteins, such that after co-digestion with trypsin and MALDI-ToF MS analysis, ten analyte:standard peptide pairs can be isolated from the mass spectrum (Figure 44). These were subsequently used for the absolute quantification of ten analyte proteins (absolute quantification for proteins achieved is discussed in the context of the biological experiment, section 3.5). In the event that previous experience of the analyte system is not feasible, it is important that the design of a quantification strategy for proteins considers variation in ion signal response which is especially inherent with MALDI-ToF MS analysis (Baumgart *et al.,* 2004). In particular, arginine terminated peptides are known to yield more abundant signals than those terminated with lysine (Krause *et al.,* 1999). In a complex MALDI-ToF mass spectrum, peptides that are abundant and have a high response factor dominate the spectrum. Tryptic Q-peptides T11, T18, T12, T4 and T17 are lysine terminated (Figure 18) and it is clear that these peptides have considerably lower signal intensity than those that are arginine terminated. The exception is T17 which contains an internal arginine residue adjacent to a proline, thus trypsin does not cleave but the presence of arginine increases the basicity of the peptide sufficiently for efficient ionisation. To achieve increased signal intensity from lysine terminated peptides, guanidination was used to convert lysine into the more basic homoarginine by reaction with *O*-methylisourea (Hale *et al.,* 2000). Guanidination of a tryptic digest of unlabelled QconCAT protein was effective at increasing the signal intensity of lysine

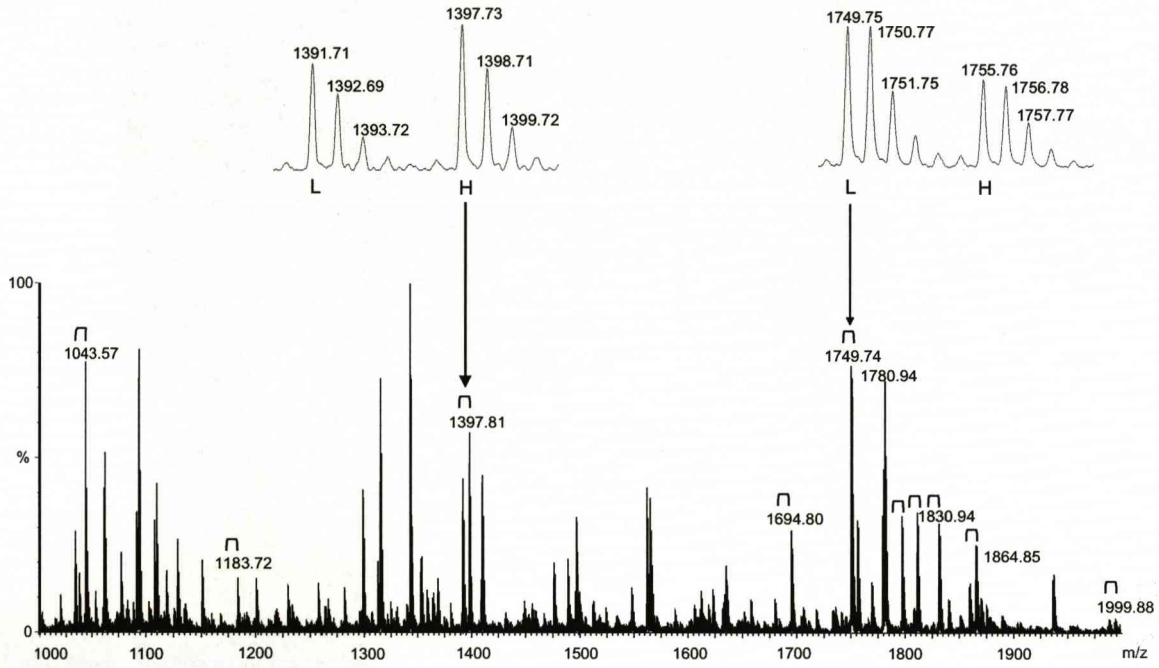



**Figure 44. Heavy:light ion pairs for analyte protein quantification in MALDI-ToF MS.** QconCAT protein was added in a 1:10 (QconCAT:chicken skeletal muscle protein) ratio to chicken skeletal muscle soluble fraction samples taken from both broiler and layer strains. The mixture was diluted 10 fold with 50mM ammonium bicarbonate, and 10% (v/v) acetonitrile was added prior to addition of trypsin (20:1 substrate:protease). The reaction mixture was incubated at 37°C for 24h after which the digest was incubated with additional trypsin (20:1 substrate:protease) to achieve complete digestion and 1µL was analysed by MALDI-ToF MS. Some analyte:standard pairs could be readily recognised by virtue of their masses and the 6Da separation provided by the terminal [$^{13}C_6$]arginine/lysine labelling, these are highlighted by Π and two of which (beta enolase, 1391.71m/z and glyceraldehyde 3-phosphate deydrogenase, 1749.75m/z) are expanded and inserted above the main spectrum.

terminated peptides in MALDI-ToF MS, with the exception of T17, with the conversion to homoarginine causing an increase in mass of 42Da (Figure 45). This increased signal intensity was expressed relative to the signal intensity of the base peak, T3 in MALDI-ToF mass spectra of guanidinated and non-guanidinated peptides for both labelled and unlabelled QconCAT protein digests (Figure 46). It is clear that the signal intensity relative to the base peak in each spectrum has significantly increased for lysine terminated peptides T4, T11 and T12, with the peak corresponding to guanidinated peptide, T18 overlapping with arginine terminated peak T8, thus complicating analyses. It is also evident from this graphical representation of signal intensity that the relative signal intensity of tryptic Q-peptide T17 is not affected by guanidination to the same extent, thus the internal arginine residue is sufficient for efficient ionisation by MALDI. Guanidination was also applied to chicken skeletal muscle soluble proteins with added QconCAT tryptic peptides and the subsequent increase in signal intensity of lysine terminated peptides was used to quantify an additional two proteins by MALDI-ToF MS during growth of muscle in both chicken strains; embryonic myosin and triose phosphate isomerase (Figure 47; absolute quantification for proteins achieved is discussed in the context of the biological experiment, section 3.5).

As an alternative to chemical modification of complex samples to improve ionisation, a greater number of peptides may be detected from a complex sample following peptide separation prior to detection by mass spectrometry. Pre-fractionation of peptides is often achieved by liquid chromatography separation through a reversed phase gradient packed into a capillary column. This separates peptides according to their hydrophobicity which is largely determined by their constituent amino acids. Reversed phase high performance liquid chromatography (RP-HPLC) of peptides is coupled directly to liquid phase, electrospray ionisation with subsequent mass spectral detection. Co-digested chicken skeletal muscle soluble proteins and QconCAT were analysed by RP-HPLC-ESI-Q-ToF MS using an EASY-nLC (Proxeon, Denmark) with separation of peptides over a 50min gradient of acetonitrile (0-100%) at a flow rate of 200nL/min. Extracted ion chromatograms demonstrated co-elution of both peptides, for example triose phosphate isomerase and beta enolase, where both QconCAT and analyte peptides eluted at 26.91min and 23.08min respectively (Figure 48). Due to the high level of sample complexity, extracted ion chromatograms contained peptides of the same m/z eluted at different times but mass spectra resulting from these peaks did not contain Q-peptides. For absolute quantification, extracted ion chromatograms for unlabelled (analyte) and labelled (QconCAT) peptides were used to locate the ions, and the chromatographic boundaries of the

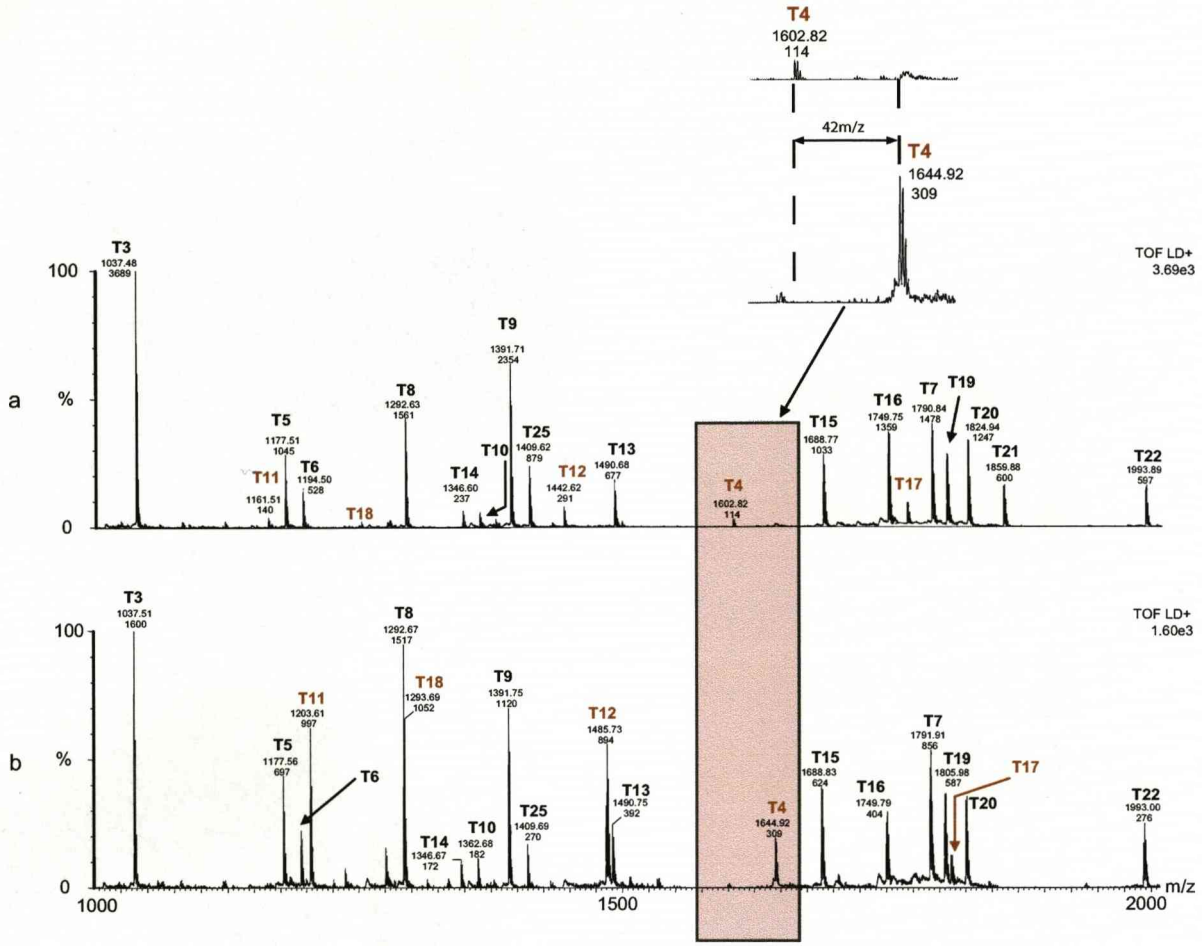



**Figure 45. Guanidination of QconCAT peptides.**
Unlabelled QconCAT protein was digested in-solution with trypsin and peptides were analysed by MALDI-ToF MS (a). To enhance the signal intensity of lysine terminated peptides in MALDI-ToF MS, lysine residues were converted to the more basic homoarginine by guanidination (Hale *et al.*, 2000). This reaction was carried out by drying down the peptide mixture and reconstituting in 10µL 7M ammonia solution to which was added 5µL 0.5M *O*-methylisourea (in $ddH_2O$). This was mixed thoroughly and incubated overnight at room temperature prior to drying down and de-salting using $C_{18}$ ZipTips (Millipore, Watford, UK). Guanidinated peptides were analysed by MALDI-ToF MS (b). Lysine terminated peptides are indicated in red.

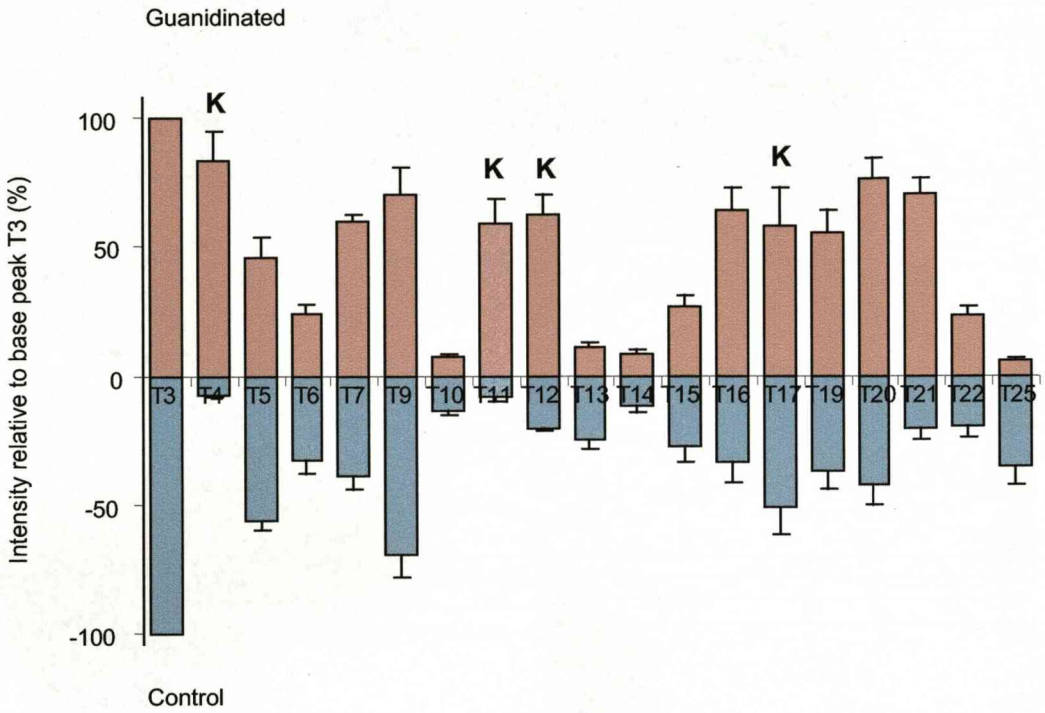



**Figure 46. Relative signal intensity of guanidinated QconCAT peptides.**
QconCAT tryptic peptides were guanidinated and analysed by MALDI-ToF MS. Signal intensity of all peptides was expressed relative to the base peak in each spectrum; T3. Signal intensity of lysine terminated peptides (K; T4, T11, T12 and T17) was compared before (blue bars) and after (pink bars) guanidination. Data are expressed as the mean±sem where n=10 (5 $^{15}$N labelled, 5 unlabelled).

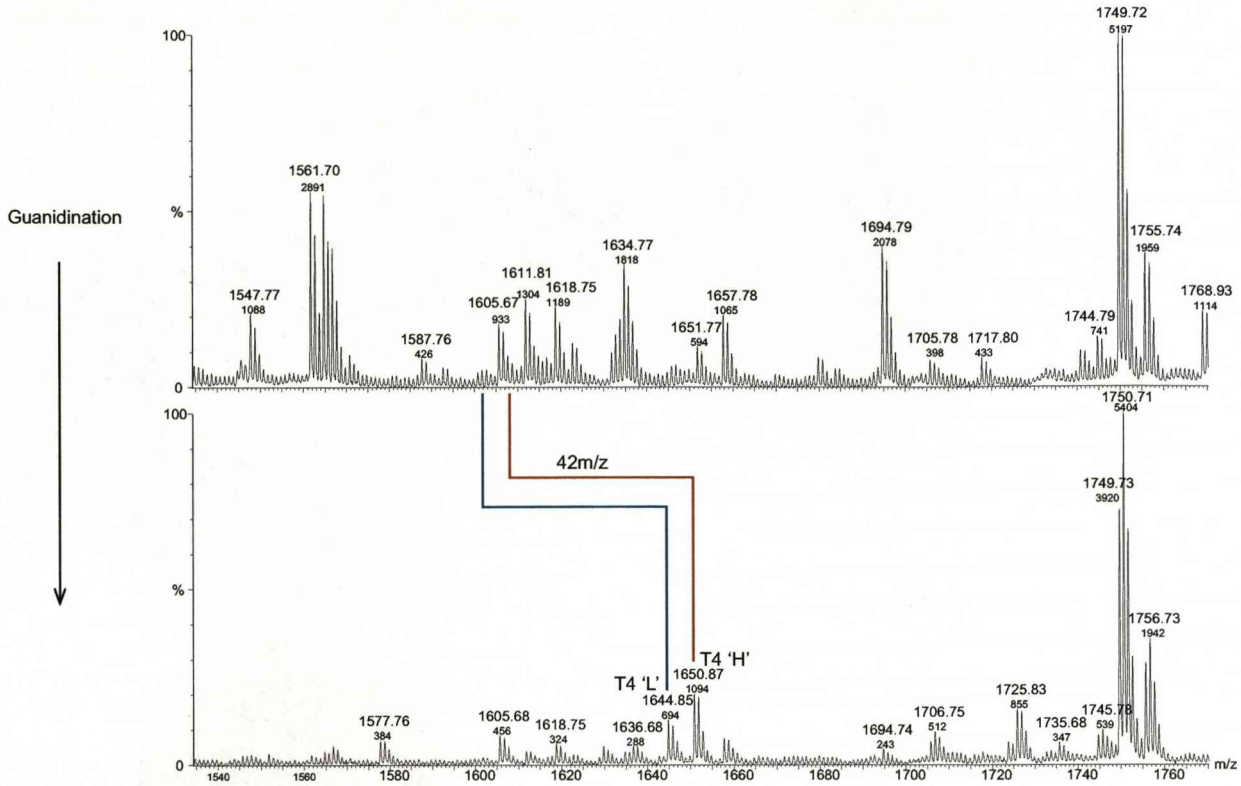



**Figure 47. Guanidination of chicken skeletal muscle soluble peptides.**
QconCAT protein (7µg) was added to each preparation of chicken skeletal muscle soluble fraction (70µg protein) of birds from 1d to 30d post hatching, with four birds at each time point. These mixtures were digested with trypsin to completion and guanidinated with *O*-methylisourea overnight. Mixtures were de-salted using $C_{18}$ ZipTips (Millipore, Watford, UK), prior to analysis by MALDI-ToF MS. Lysine terminated peptides that were guanidinated were evident from a mass shift of 42Da. This is illustrated for the lysine terminated QconCAT peptide representing triose phosphate isomerase (before guanidination; $[M+H]^+$ 'L' 1602.87, $[M+H]^+$ 'H' 1608.87m/z).

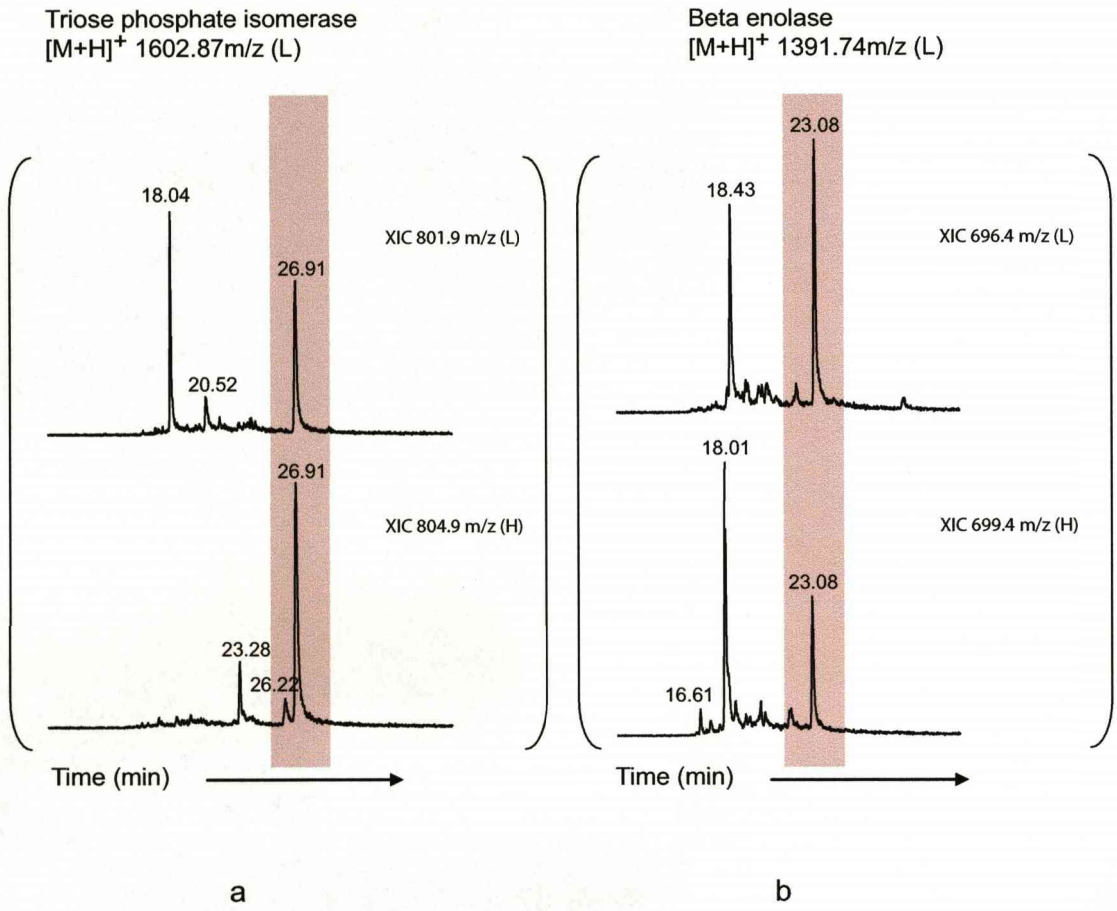



**Figure 48. Chromatographic elution of analyte and internal standard peptides.**
QconCAT protein was added in a 1:10 (QconCAT:chicken skeletal muscle protein) ratio to chicken skeletal muscle soluble fraction samples taken from both broiler and layer strains. The mixture was diluted 10 fold with 50mM ammonium bicarbonate, and 10% (v/v) acetonitrile was added prior to addition of trypsin (20:1 substrate:protease). The reaction mixture was incubated at 37°C for 24h after which the digest was incubated with additional trypsin (20:1 substrate:protease) to achieve complete digestion. Peptide mixtures were analysed by LC-ESI-Q-ToF MS. Extracted ion chromatograms were performed for both unlabelled analyte and labelled internal standard peptides to confirm elution times. For triose phosphate isomerase, 'L' and 'H' Q-peptides eluted at 26.91min and for beta enolase, the peptide pair for absolute quantification co-eluted at 23.08min. This was confirmed by combining the entire area under each peak to produce a mass spectrum. In both instances, extracted ion chromatograms revealed alternative species of the same mass eluting at different times but the resulting mass spectrum confirmed that these chromatographic peaks did not contain Q-peptides.

coincident pair of peptides were used to delineate the combined mass spectra, from which peptides were quantified by mass spectrometric intensities of the doubly charged ions; there was no evidence of higher charge states, for example $[M+3H]^{3+}$ corresponding to analyte:QconCAT pairs (Figure 49). Correlation between quantification achieved by this method and by MALDI-ToF MS was excellent and has been discussed previously (Figure 42, section 3.3.1).

For some aspects of quantitative proteomics, MALDI-ToF MS has advantages. Data can be accumulated for a variable number of laser shots, ensuring comparable signal intensities between replicates. Virtually all of the signal resides in the singly charged $[M+H]^{+}$ ion, whereas with electrospray ionisation, the signal can be distributed over a number of differently charged species (although often, as for analysis of trypsin digested chicken skeletal muscle soluble proteins, peptides used for absolute quantification were doubly charged; $[M+2H]^{2+}$). However, for complex analytical mixtures, the density of a MALDI-ToF spectrum, coupled with a noisy baseline, can compromise quantification. One approach to simplification of MALDI-ToF MS analyses relies on prior fractionation of the peptide mixture before deposition of successive fractions on the MALDI-ToF target (Mirgorodskaya *et al.,* 2005). This technique was investigated for unlabelled QconCAT peptides and a mixture of [$^{15}$N] labelled QconCAT and unlabelled QconCAT to test its application and ability to separate QconCAT peptides. Peptides separated over a 90 minute gradient of acetonitrile (0-100%) were collected onto a MALDI target in one minute fractions from 42 to 52 minutes (Figure 50). This gave excellent separation and confirmed that QconCAT labelled and unlabelled peptides behave in the same way when separated by reversed phase; both eluted at the same time as [$^{15}$N] does not separate from [$^{14}$N] by reversed phase, and signal intensity ratios of unlabelled to labelled peptides remained constant throughout the fractions collected. For the analysis and quantification of chicken skeletal muscle soluble proteins, co-digested analyte and standard proteins were separated by reversed-phase chromatography with fractions (200nL) collected at one minute intervals onto a MALDI target for analysis by MALDI-ToF MS (Figure 51). This provided an efficient detection system with peptides fixed in the solid phase for continued interrogation when acquiring data for quantification. LC-MALDI-ToF MS was used for analysis of a single chicken skeletal muscle sample (30d layer) to highlight the potential benefit of this method. This approach allowed quantification of all proteins selected for incorporation into the QconCAT protein and present in the analyte system, contributing additional information for quantification. Comparing quantification by LC-MALDI-ToF MS with MALDI-ToF MS for ten proteins identified by both

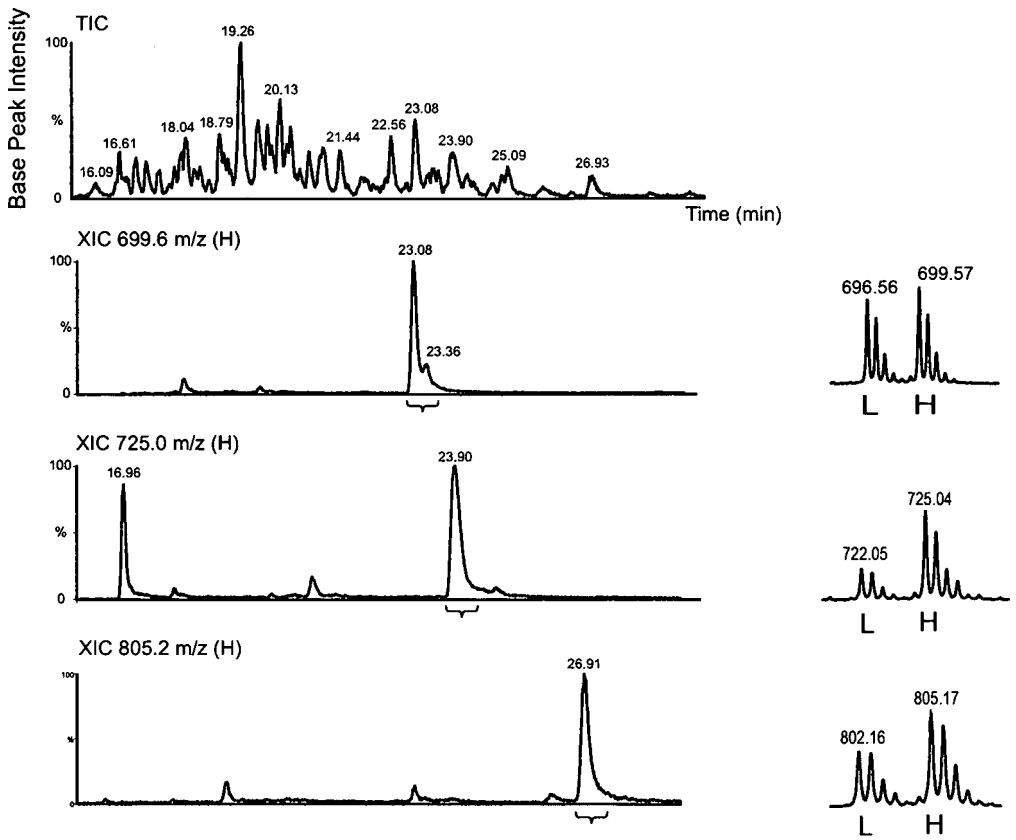



**Figure 49. Isolation of analyte:standard peptide pairs by reversed phase chromatography.** QconCAT protein (7µg) was added to chicken skeletal muscle soluble fraction (70µg protein). This mixture was digested with trypsin and analysed by LC-ESI-Q-ToF MS. All peptide pairs for quantification were present as doubly charged ions; there was no evidence of triply charged species. The upper panel is the total ion chromatogram (base peak intensity) of the elution profile from 16 to 29min. The lower panels are the extracted ion chromatograms for representative QconCAT peptides of doubly charged ions (beta enolase, 699.6m/z, eluted at 23.08min, glycogen phosphorylase, 725.0m/z, eluted at 23.90min and triose phosphate isomerase, 805.2m/z, eluted at 26.91min) with corresponding mass spectra showing analyte and QconCAT peptide ion pairs used for quantification presented as inserts on the right.

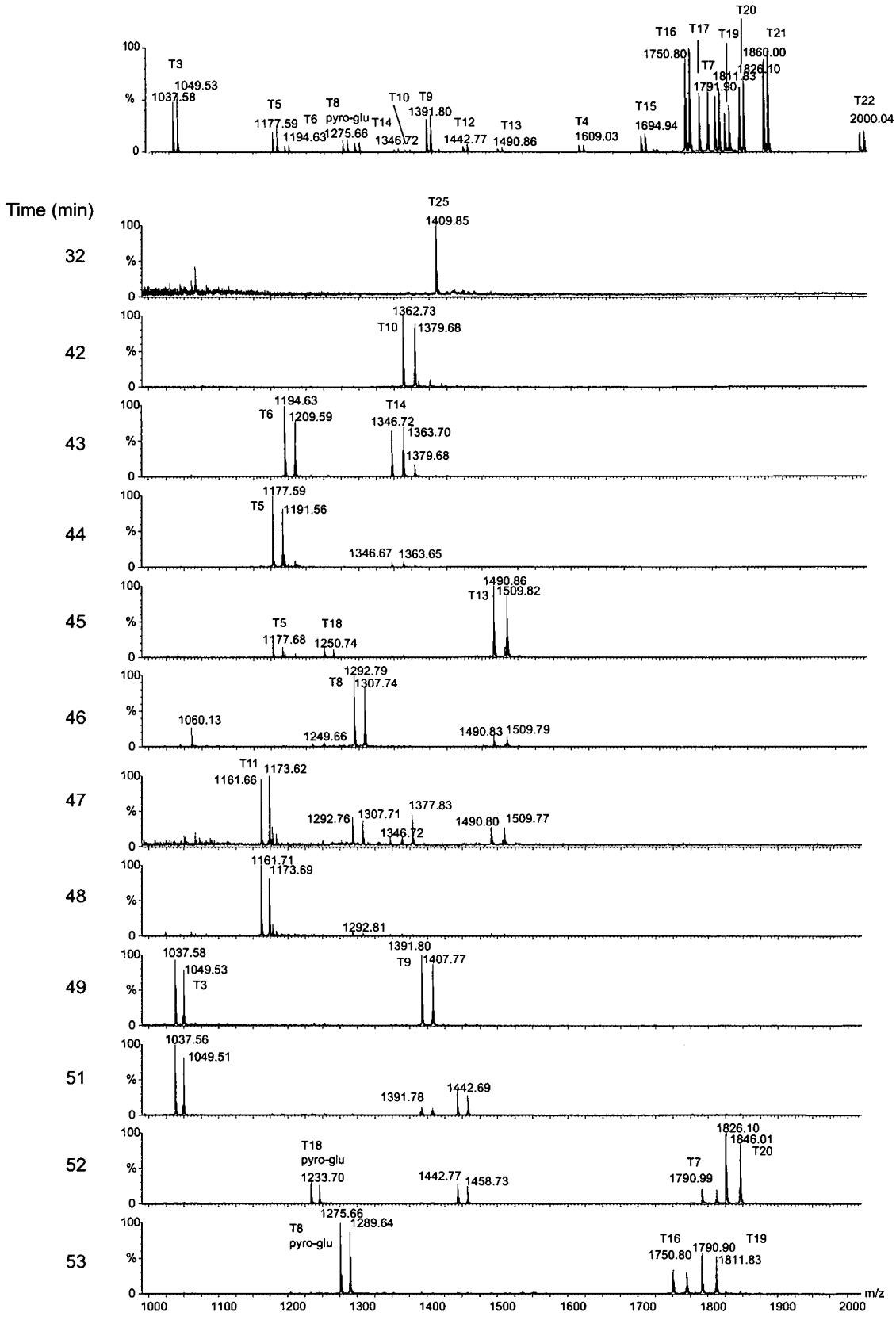



**Figure 50. Fractionation of QconCAT peptides by LC-MALDI-ToF MS.**
QconCAT proteins both unlabelled and [$^{15}$N] labelled were mixed in an approximate 1:1 ratio and digested with trypsin. Peptides were analysed by MALDI-ToF MS (top spectrum) and were separated by reversed phase HPLC (EASY-nLC, Proxeon, Denmark) over a 60min gradient of acetonitrile (0-100%) at a flow rate of 200nL/min. Fractions were collected manually onto a MALDI target every minute and analysed by MALDI-ToF MS. MALDI-ToF mass spectra of fractions containing the majority of eluted QconCAT peptides are shown.

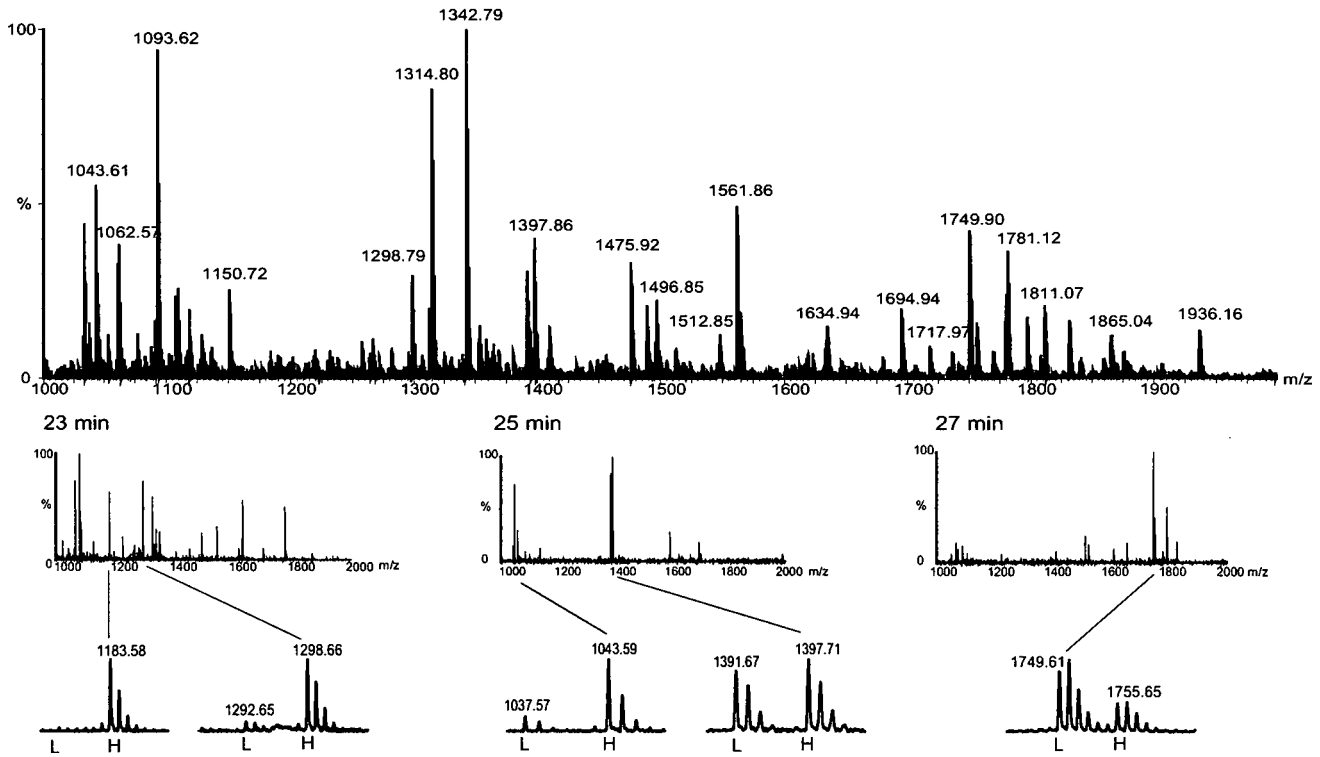



**Figure 51. Quantification of chicken skeletal muscle soluble proteins by LC-MALDI-ToF MS.**

QconCAT protein (7µg) was added to a preparation of chicken skeletal muscle soluble fraction (70µg protein) in a ratio of 1:10. This mixture was digested with trypsin and analysed by LC-MALDI-ToF MS. Peptides were separated over a 50min reversed phase acetonitrile gradient (0-100%) and fractions of 1min (200nL) were collected directly onto a Waters MALDI-ToF target. The upper panel is the MALDI-ToF mass spectrum of the entire digest, the lower panels illustrate three fractions collected from the reversed phase eluate at 23, 25 and 27 min. Representative pairs of analyte: standard peptides are highlighted.

methods (five of which were quantified as 0nmol/g by both methods) confirms that both methods of analysis give consistent and comparable quantification (Figure 52).

## 5.4 Validation of the QconCAT Method

### 5.4.1 Quantification of unlabelled QconCAT by labelled QconCAT

To use the QconCAT protein as an internal standard for the absolute quantification of chicken skeletal muscle soluble proteins by co-digestion and MS, it is essential that QconCAT added produces peptides in equivalent amounts. The QconCAT was designed to achieve this, but to confirm, [$^{13}C_6$]arg/[$^{13}C_6$]lys-labelled and unlabelled QconCAT proteins were mixed in known ratios from 0 to 1.2 (L:H) and co-digested with trypsin. The resulting signal intensity ratio of each peptide pair upon MALDI-ToF MS analysis was detected and the correlation between protein ratio added and peptide ratio measured was good with a slope of 0.99 and a correlation ($R^2$) of 0.997 (Figure 53). This reinforced that absolute quantification can be achieved by adding a known amount of QconCAT protein and basing quantification of analyte proteins on the signal intensity of each internal standard peptide.

### 5.4.2 Variance in the QconCAT method

As an assessment of variance due to the analytical procedure, four identical protein mixtures (70µg chicken skeletal muscle with 7µg QconCAT) were digested with trypsin and the surrogate peptides were used to quantify proteins by MALDI-ToF MS. Quantification data were collected and used to assess analytical variance (Figure 54a). The reproducibility of the method was high, and in both instances, the analytical variance was significantly lower than that for quantification measured for four different birds of each strain (Figure 54b). For example, the analytical variance (CV 6.0%) for β eno, n=4) compared favourably to biological variance (CV 24.0%, β eno, n=4). Increasing the number of analytical replicates to 10 had very little effect on analytical variance (CV 6.0% β eno, n=10; Figure 55). To assess the extent to which this is affected by the signal intensity of each peptide, particularly for MALDI-ToF MS, in which peptides that do not ionise efficiently are often difficult to distinguish from the background noise, the coefficient of variation for analytical replicates was compared with the signal intensity of each analyte peptide. For five proteins of varying signal:noise ratio in a typical mass spectrum of analyte and internal standard peptides, for four analytical replicates, the signal intensity of each analyte peptide was expressed as a proportion of the total intensity for the five analyte peptides. This was compared with the coefficient of variation measured for the same

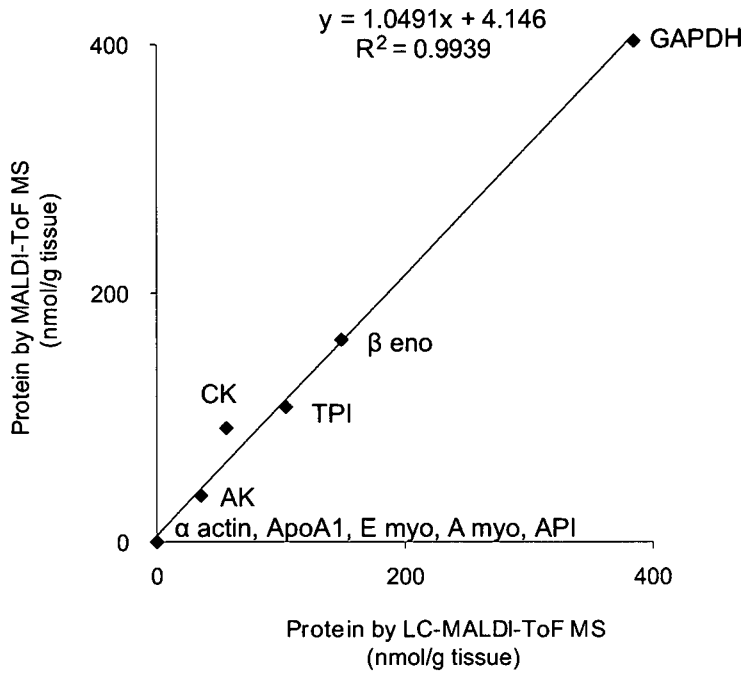



**Figure 52. Consistency of protein quantification using LC-MALDI-ToF MS and MALDI-ToF MS.**

QconCAT protein (7µg) was added to a preparation of chicken skeletal muscle soluble fraction (70µg) in a ratio of 1:10. This mixture was digested with trypsin and analysed by LC-MALDI-ToF MS or MALDI-ToF MS. For a subset of proteins in a single biological sample, quantification acquired through different analytical modalities was compared. Five proteins were quantified as 0nmol/g by both methods (α actin, ApoA1, E myo and API).

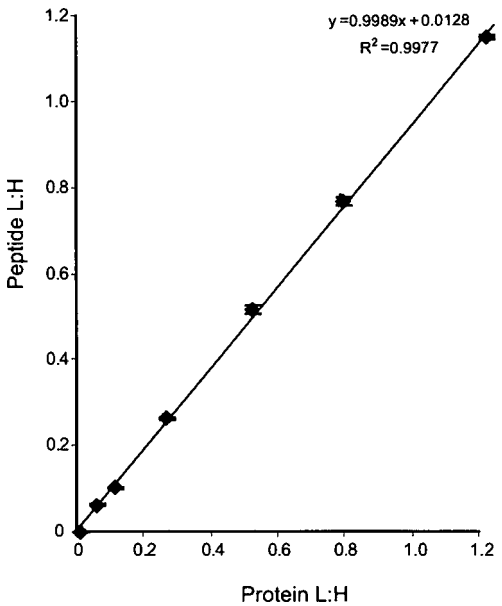



**Figure 53. Validation of quantification using a mixture of unlabelled and labelled QconCAT proteins**

QconCAT unlabelled (L) and labelled (H) proteins were mixed in increasing ratios from 0 to 1.2 (L:H) and digested in-solution with trypsin at a ratio of protein:trypsin of 20:1. Peptides were analysed by MALDI-ToF MS and the relative signal intensity of unlabelled and labelled peptide ions was used to calculate peptide L:H ratio. Data are presented as mean±sem (n=17).

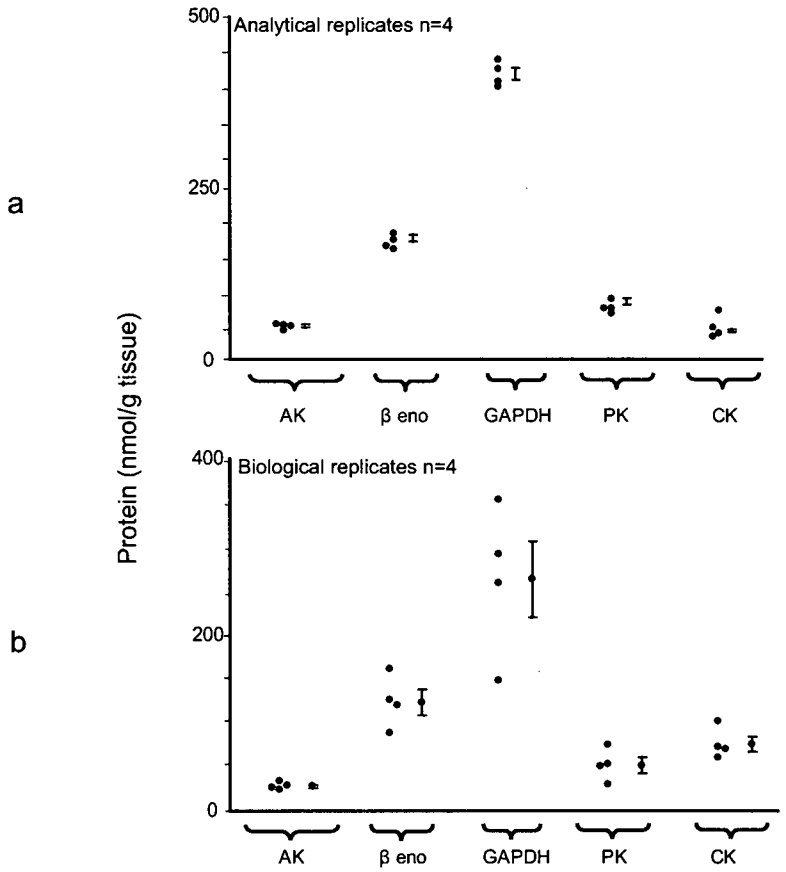



**Figure 54. Sources of variance in a QconCAT experiment.**
Soluble protein from chicken pectoralis muscle (70µg) was mixed with QconCAT protein (7µg) in four technically replicated experiments and digested to completion with trypsin. For each protein, individual data points are plotted to the left of mean±sem for the same bird where n=4 (a) and for four different birds to demonstrate biological variance (b).

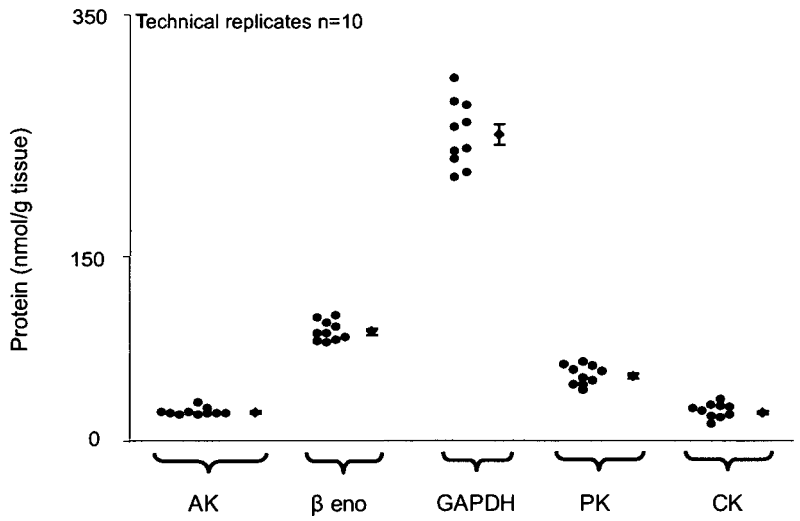



**Figure 55. Sources of variance in a QconCAT experiment.**
Soluble protein from chicken pectoralis muscle (70µg) was mixed with QconCAT protein (7µg) in 10 technically replicated experiments and digested to completion with trypsin. For each protein, individual data points are plotted to the left of mean±sem for the same bird where n=10.

analyte peptides when quantified using QconCAT, as an assessment of analytical variance (Figure 56). As expected, coefficient of variation is greater for peptides with low signal intensity, suggesting that signal:noise ratio of each peptide pair used for quantification has a significant impact on the reproducibility of absolute quantification. In this biological study, the criterion for using a peptide ion for absolute quantification was user discretion, but these results demonstrate that it may be necessary to apply a threshold of signal intensity, below which peptide ions are not used for quantification in order to reduce analytical variation. However, variation between repeated analyses of the same sample, reflecting the nature of mass spectrometry in that each ionisation may vary slightly with time, plate position, presence of matrix crystals and other experimental conditions, is so low compared to variation between individual animals that the effects are negligible in the context of this biological study.

### 5.4.3 Accuracy of the QconCAT method

To assess the accuracy of a QconCAT experiment for quantification, a known amount of purified proteins adenylate kinase (AK) and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) from chicken skeletal muscle were independently mixed with a known amount of [$^{13}C_6$]arg/[$^{13}C_6$]lys-labelled QconCAT protein and co-digested with trypsin. Signal intensities of analyte peptides for these proteins represented in the QconCAT, and QconCAT peptides, were used for absolute quantification. The amount of protein added was compared with amount of protein measured using QconCAT, giving excellent, linear correlation ($R^2$=0.99 for both proteins, Figure 57). To assess accuracy within the analytical environment, a known amount of AK was spiked into chicken skeletal muscle soluble fraction from a 30d broiler. The amount of AK added was converted into protein concentration as nmol/g tissue and compared with the total concentration of AK in the tissue (nmol/g) as quantified using QconCAT (Figure 58). As expected, there was a strong correlation ($R^2$=0.9992) with a slope of 1, indicating the lack of any systematic quenching effects over an extended dynamic range.

### 5.4.4 Comparison of the QconCAT method with alternative strategies for absolute quantification

Although there is nothing formally different between a chemically synthesised peptide and a peptide excised from a QconCAT by proteolysis, quantification was compared using the two methods. The synthetic peptide, of sequence LVSWYDNEFGYSNR and mass 1748.77Da representing the abundant protein GAPDH, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset

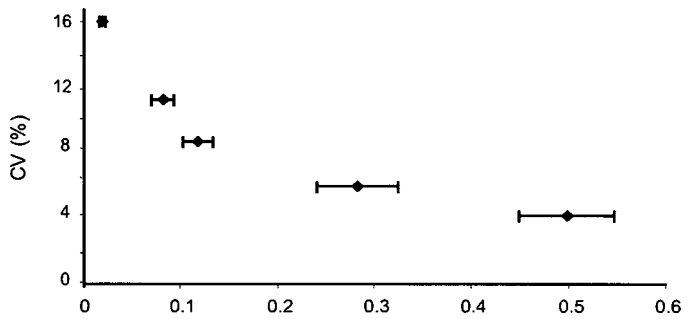



**Figure 56. Relationship between coefficient of variance and ion signal intensity in MALDI-ToF mass spectra.**
Soluble protein from chicken pectoralis muscle (70µg) was mixed with QconCAT protein (7µg) in four technically replicated experiments and digested to completion with trypsin. For five proteins, signal intensity of analyte peptides is expressed as a ratio of the total ion count for the five analyte peptides used in each spectrum and plotted against coefficient of variance for quantification of that particular protein by QconCAT. Signal intensity data are plotted as mean±sem for the same bird where n=4.

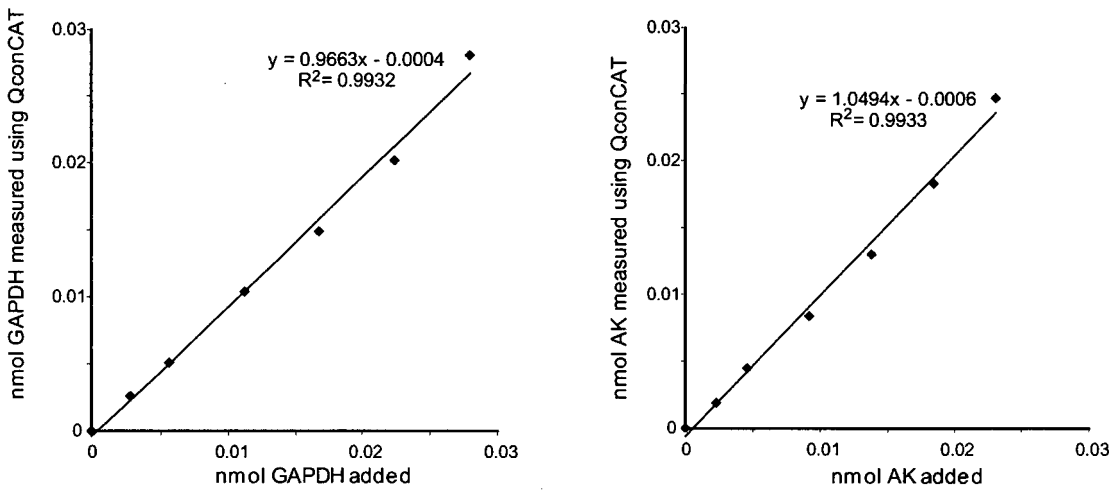



**Figure 57. Accuracy of the QconCAT method using purified proteins.**
Purified glyceraldehyde 3-phosphate dehydrogenase (GAPDH; a) and adenylate kinase (AK; b) were mixed in known amounts from 0.1µg to 1.0µg with labelled QconCAT protein. Protein mixtures were digested in-solution with trypsin at a ratio of total protein:enzyme of 20:1. Relative signal intensity of analyte and internal standard peptide ions was used to quantify the amount of each protein.

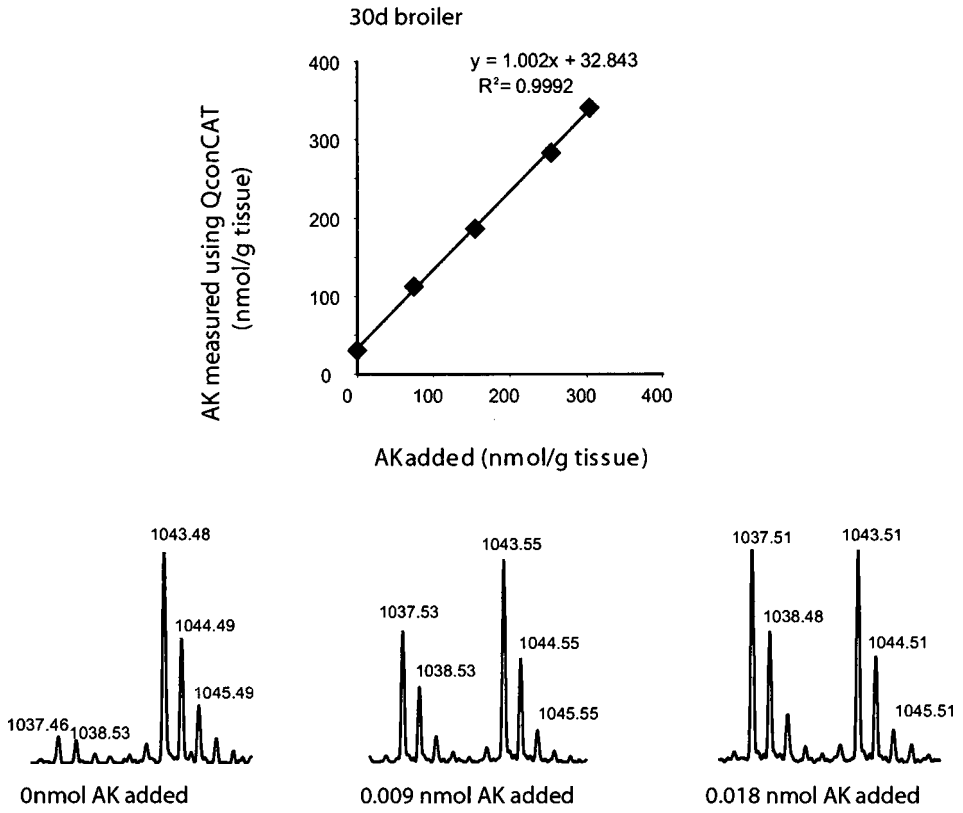



**Figure 58. Accuracy of quantification using QconCAT.**
Purified adenylate kinase (AK) was added to chicken skeletal muscle soluble fraction from a 30d broiler. AK was added from 0nmol to 0.02nmol which resulted in a final protein concentration of 0nmol/g to 300nmol/g and the amount of AK in the tissue was quantified by adding 0.015nmol QconCAT prior to digestion with trypsin. Proteolysis was allowed to continue for 24h after which peptides were analysed by MALDI-ToF MS. The upper panel shows the correlation between AK added and that quantified in the muscle using QconCAT after digestion with trypsin. Spectra showing the change in MALDI-ToF mass spectral signal intensity over the range of protein concentrations used in this experiment are shown beneath.

from the analyte peptide. Initially, a known amount of unlabelled QconCAT was added to a known amount of synthetic peptide to compare quantification. QconCAT was used to quantify the amount of synthetic peptide (Figure 59). Although correlation was good ($R^2$=0.999), there was a consistent under-estimate of the amount of synthetic peptide as calculated by QconCAT. This is likely to reflect the way in which the two standards are themselves independently quantified but to investigate further, preparations of synthetic peptide were stored under different conditions for 24h prior to quantification by QconCAT (Figure 60). Storage in the original vial at 4°C in sterile water produced consistent quantification with good correlation and the same under-estimate of synthetic peptide amount when quantified in this way. However, aliquots of synthetic peptide in 50mM ammonium bicarbonate stored at -20°C in Eppendorf tubes produced variable quantification with little correlation, or agreement, even between two preparations stored under identical conditions. This suggests that the synthetic peptide becomes degraded with freeze-thawing and also that it may stick to the surface of the Eppendorf tube. Future preparations of synthetic peptide were maintained in their original vial and used upon reconstitution for the most accurate results. By contrast, QconCAT protein is stored at -20°C in Eppendorf tubes and consistently delivers the same quantification; this gives a distinct advantage over the synthetic peptide approach. Using freshly prepared synthetic peptide, quantification of a single protein (GAPDH, which exhibits a dramatic change in abundance during post hatching development) using the QconCAT-derived peptide and the identical synthetic peptide was compared. Both the synthetic peptide and QconCAT protein were added to the same analytical sample such that the three peptides were easily distinguished in the mass spectrum. The correlation between data obtained using QconCAT and that obtained using the synthetic peptide was high ($R^2$=0.998) (Figure 61), and quantification data were consistent using either internal standard. A small consistent discrepancy (less than 10%) between the two methods could be attributable to the method of quantification used for the two standards. The discrepancy between the synthetic peptide and the QconCAT was reduced if the latter was used to quantify the former, but was still present. This residual discrepancy is difficult to explain but is not attributable to incomplete digestion of the QconCAT (Figure 33, section 3.2.1). In the case of the QconCAT, a protein assay was used to determine the amount of protein, as this was the same method used to quantify total protein in the analyte. For the synthetic peptide, the quantity supplied by the manufacturer is too small for independent quantification, and it was necessary to assume that the quantity in the vial was indeed that specified by the manufacturer. The difference between the two standards was minor compared to the biological variance within the system, would not contribute significant

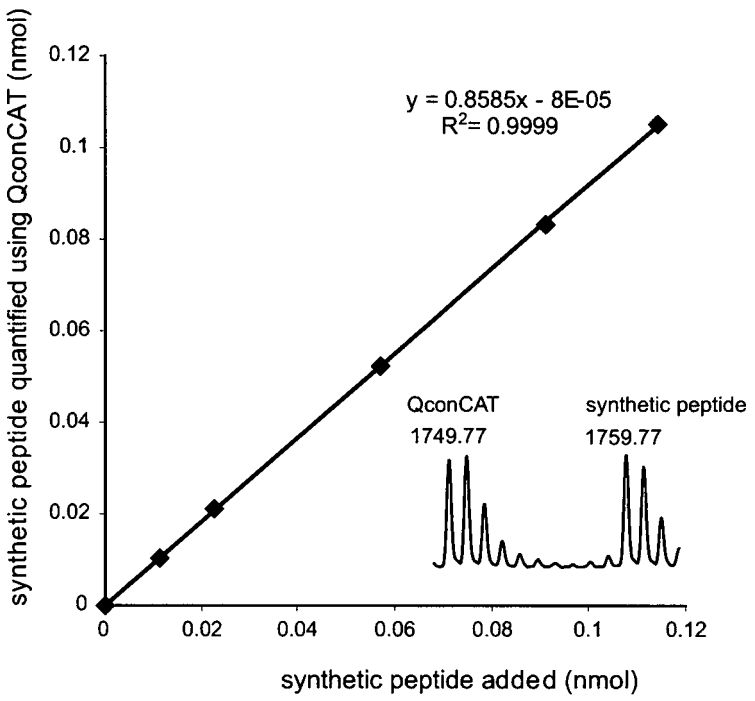



**Figure 59. Quantifcation of a synthetic peptide internal standard using QconCAT.**
A peptide, of sequence LVSWYDNEFGYSNR and mass 1748.77Da representing the abundant protein GAPDH in the QconCAT protein for quantification of chicken skeletal muscle soluble proteins, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide. 1nmol synthetic peptide was reconstituted in 1mL sterile $H_2O$, resulting in a final concentration of 1.76µg/µL. Increasing amounts of synthetic peptide (0.00-0.11nmol) were added to 0.05nmol unlabelled QconCAT protein which was digested over 24h with trypsin. Peptides were analysed using MALDI-ToF MS (see inserted spectrum) and the relative signal intensity of QconCAT and synthetic peptide ions was used for quantification.

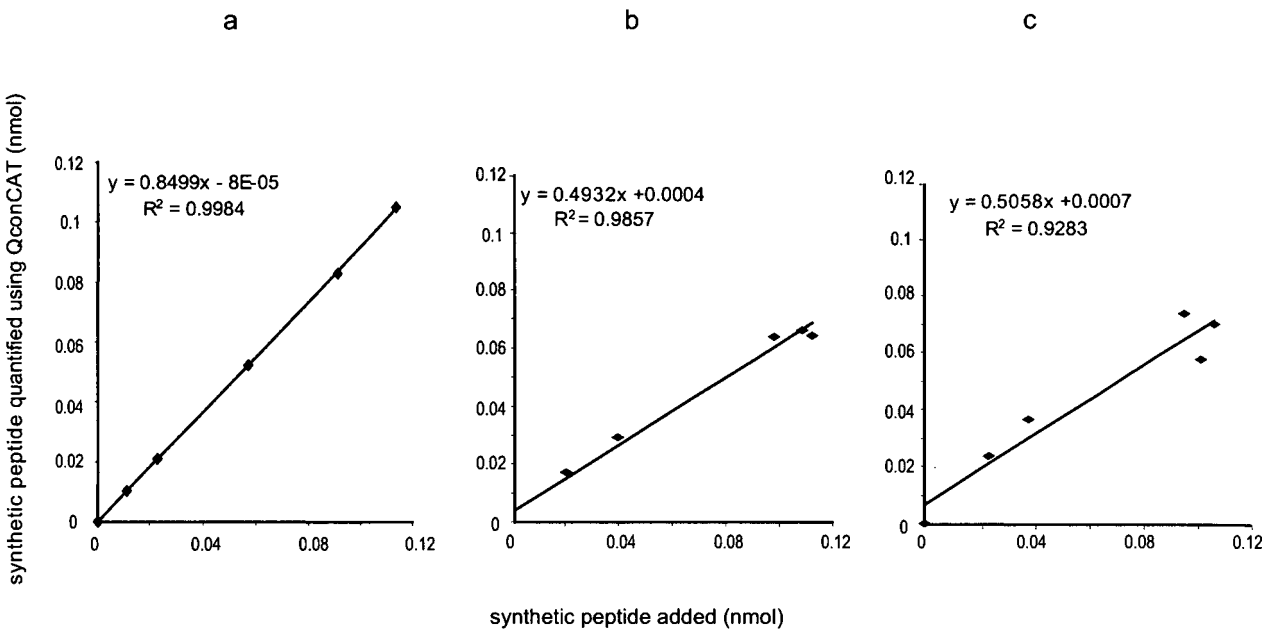



**Figure 60. Quantification of a synthetic peptide internal standard stored under different conditions, using QconCAT.**

A peptide, of sequence LVSWYDNEFGYSNR and mass 1748.77Da representing the abundant protein GAPDH in the QconCAT protein for quantification of chicken skeletal muscle soluble proteins, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide. 1nmol synthetic peptide was reconstituted in 1mL solvent, resulting in a final concentration of 1.76µg/µL. Increasing amounts of synthetic peptide (0.00-0.11nmol) were added to 0.05nmol unlabelled QconCAT protein which was digested over 24h with trypsin. Peptides were analysed using MALDI-ToF MS and the relative signal intensity of QconCAT and synthetic peptide ions was used for quantification. For comparison, synthetic peptide was reconstitued in 1mL sterile water and used immediately (a), or in 50mM ammonium bicarbonate and stored at -20°C prior to use in two separate vials (b&c).

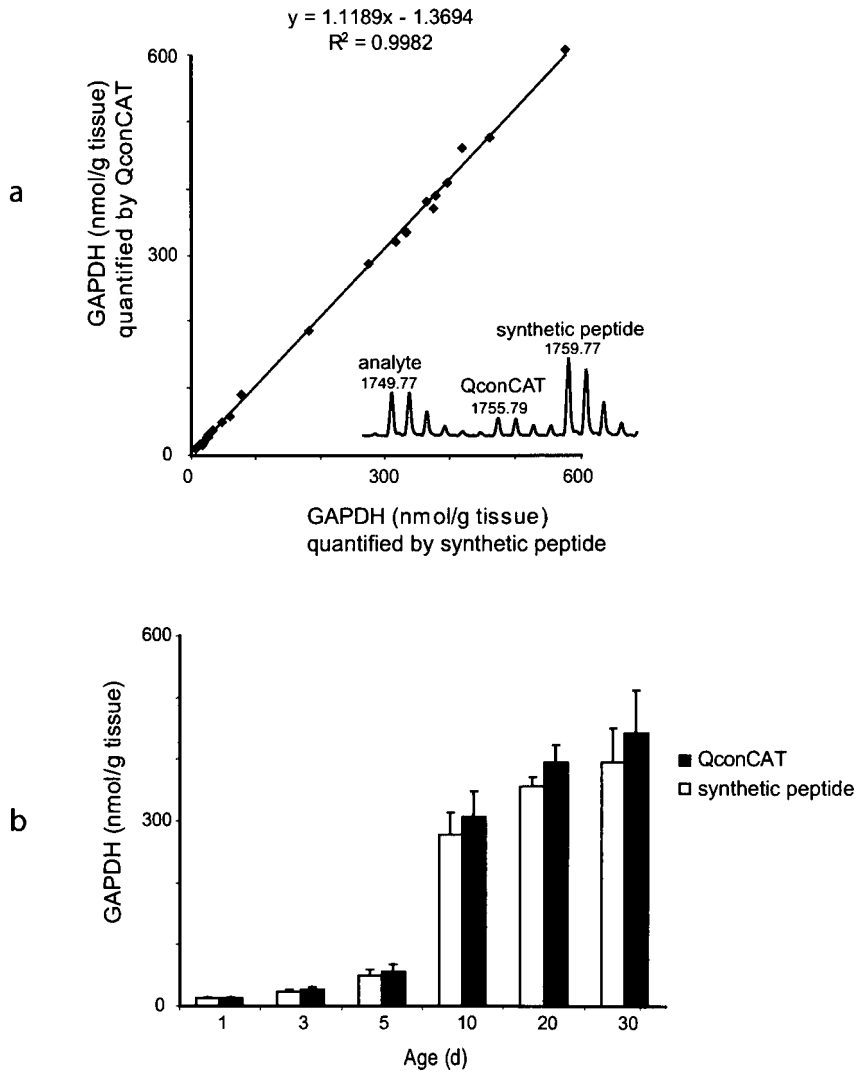



**Figure 61. Comparison of QconCAT and synthetic peptide for quantification.**
For one protein (glyceraldehyde 3-phosphate dehydrogenase; GAPDH), quantification was achieved relative to a QconCAT peptide and the same peptide chemically synthesised. For both methods, the internal standards (2µg QconCAT protein, or 0.05µg synthetic peptide) were added to 20µg chicken skeletal muscle soluble protein prior to digestion with trypsin and data were acquired using MALDI-ToF MS (see inserted spectrum). Quantification by either method was directly compared (a). The time dependent developmental expansion of GAPDH (nmol/g tissue, mean±sem, n=4) in broiler was monitored by QconCAT or synthetic peptide (b).

errors, and would be readily controlled by alternative QconCAT quantification strategies. This could include incorporation into each QconCAT protein, a common peptide for internal standard quantification by a synthetic peptide that can be labelled or unlabelled, depending on the labelling status of the QconCAT protein. This peptide, chosen for its ionisation efficiency in the detection system of choice could be used to quantify and normalise all QconCAT data to an absolute standard that is common to each one.

Quantification of selected muscle proteins by the QconCAT strategy was also compared with densitometric quantification from 1D SDS-PAGE for both broilers and layers (densitometry data supplied by I.J. Edwards) for six individual proteins (Figure 62 & 63). Comparisons with some proteins, for example enolase for broilers and layers were consistent, while for some, for example GAPDH, changes in protein expression were not well matched using the two techniques. Overall comparison of proteins quantified using both methods was not favourable ($R^2$=0.382; Figure 64). This highlights the difficulty in drawing quantitative data from 1D SDS-PAGE as many proteins overlap on the gel, therefore total intensity of a single band will not necessarily be a true reflection of a single protein, but more likely multiple proteins. To investigate the linearity of stain intensity achieved with a 1D gel, the amount of total protein loaded was increased from 0-15µg chicken skeletal muscle soluble proteins and the response was measured for five proteins using densitometry of the gel image. The amount of total protein loaded was plotted against band volume measured using densitometry with strong correlation (average $R^2$=0.959 discounting AK for which stain intensity for lower protein loading was insufficient; Figure 65a). The same amount of total protein (0-15 µg) was quantified using QconCAT, with similar overall correlation (average $R^2$=0.988) for quantification of the same five proteins (Figure 65b). This demonstrates that stain intensity when measured by densitometry from a 1D SDS-PAGE gel image is linear, and only marginally less reproducible than QconCAT for quantification of an increasing amount of known proteins, providing the stain intensity is within sufficient dynamic range for reliable detection. For the individual proteins beta enolase (β-eno) and pyruvate kinase (PK), quantification was compared directly for increased protein loaded onto a 1D gel (Figure 66). To convert stain intensity into nmol/g, the proportion of abundance contributed by each protein was extrapolated to the total amount loaded onto each lane of the gel. As previously discussed, correlation was strong using both techniques, although the amount of protein was consistently over-estimated using densitometry. Due to the high level of protein complexity exhibited here, this is likely due to other proteins of lower abundance contributing to the stain intensity at each of these positions on the gel.

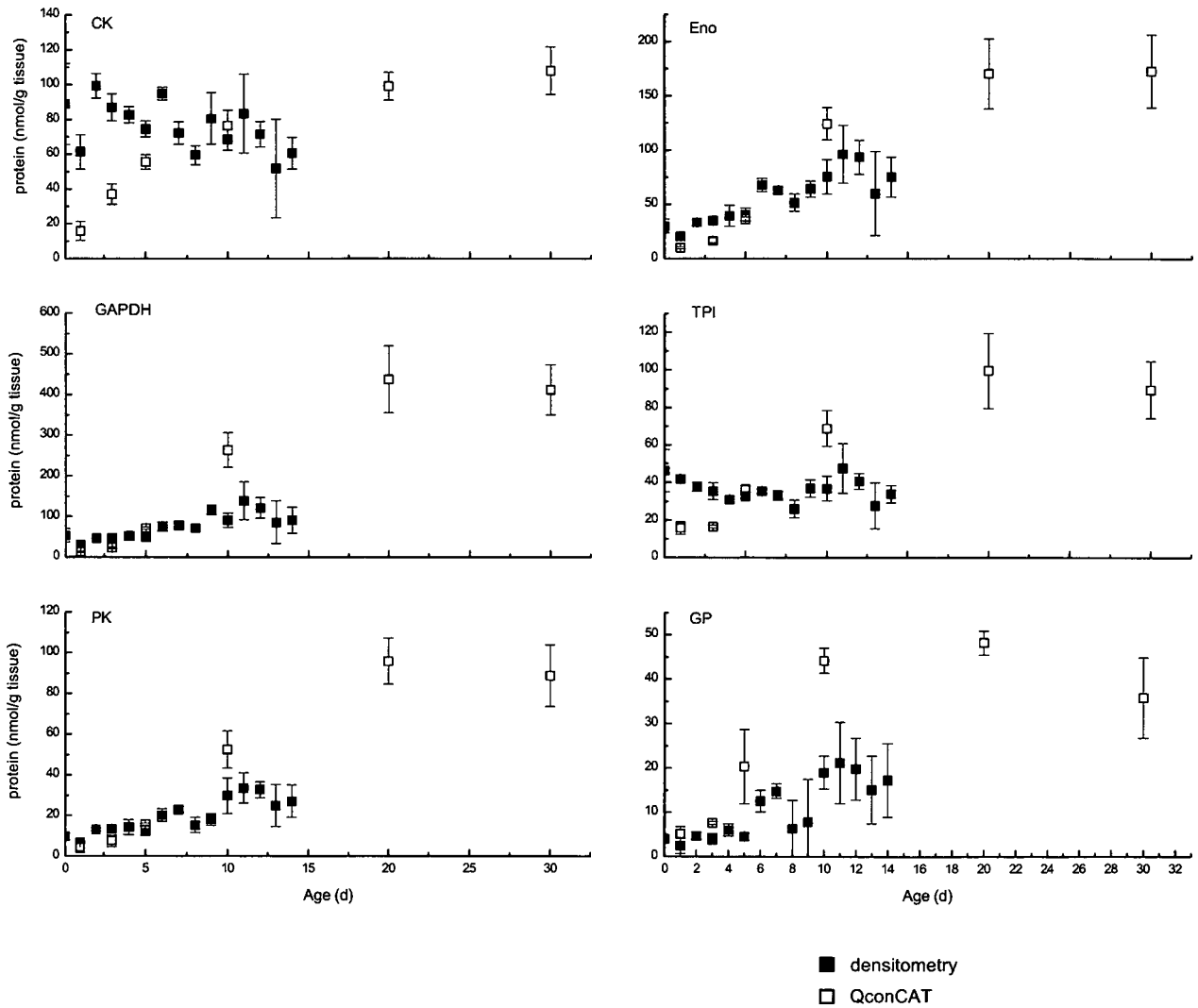



**Figure 62. Comparison of protein quantification by densitometry and QconCAT.**

For six proteins, quantification (nmol/g tissue) was compared using methods of 1D SDS-PAGE and densitometry analysis (closed squares) with QconCAT co-digestion and analysis by MALDI-ToF MS (open squares) for broiler chickens during growth (densitometry data supplied by I.J. Edwards).

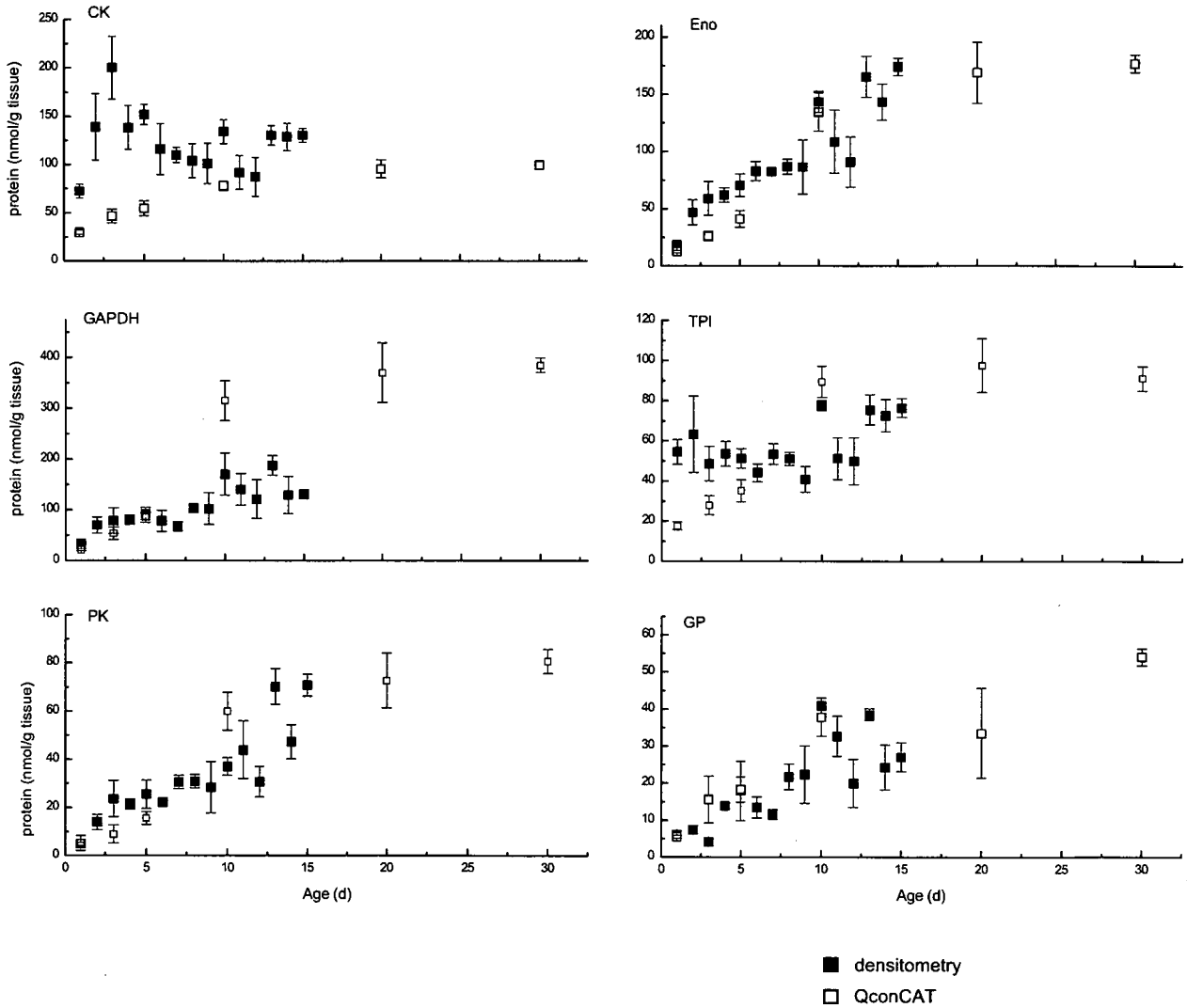



**Figure 63. Comparison of protein quantification by densitometry and QconCAT.**

For six proteins, quantification (nmol/g tissue) was compared using methods of 1D SDS-PAGE and densitometry analysis (closed squares) with QconCAT co-digestion and analysis by MALDI-ToF MS (open squares) for layer chickens during growth (densitometry data supplied by I.J. Edwards).

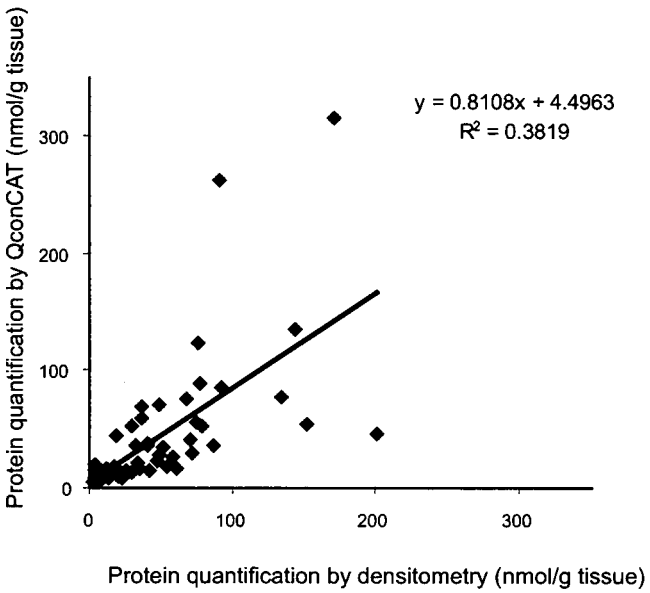



**Figure 64. Comparison of protein quantification by densitometry and QconCAT.**
For six proteins, quantification (nmol/g tissue) was compared using methods of 1D SDS-PAGE and densitometry analysis with QconCAT co-digestion and analysis by MALDI-ToF MS for layer and broiler chickens during growth (densitometry data supplied by I.J. Edwards).

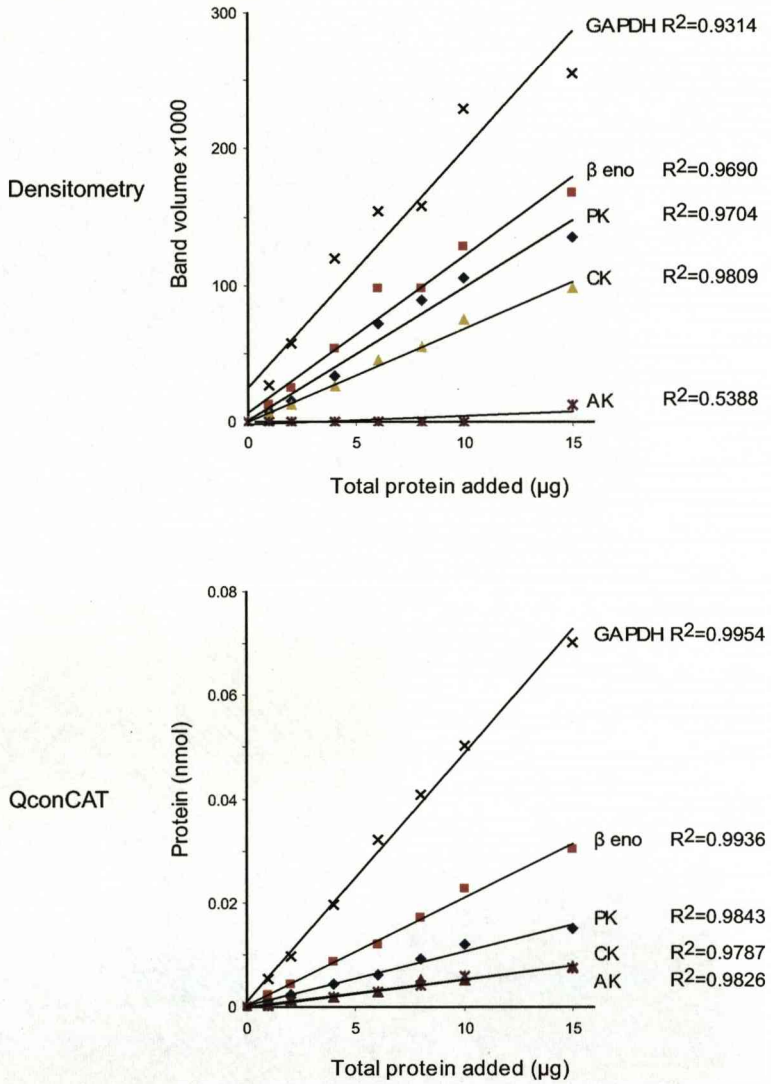



**Figure 65. Quantification of increasing amount of total protein, by densitometry and the QconCAT method.**

Chicken skeletal muscle soluble protein was analysed by SDS-PAGE with increasing amounts of total protein (µg) loaded onto the gel. The response was measured using densitometry for five proteins where the band volume was correlated with the amount of total protein. For each protein, the correlation coefficient ($R^2$) is expressed to the right of the graph. The same amount of total protein was also analysed by the QconCAT method with peptide analysis by MALDI-ToF MS.

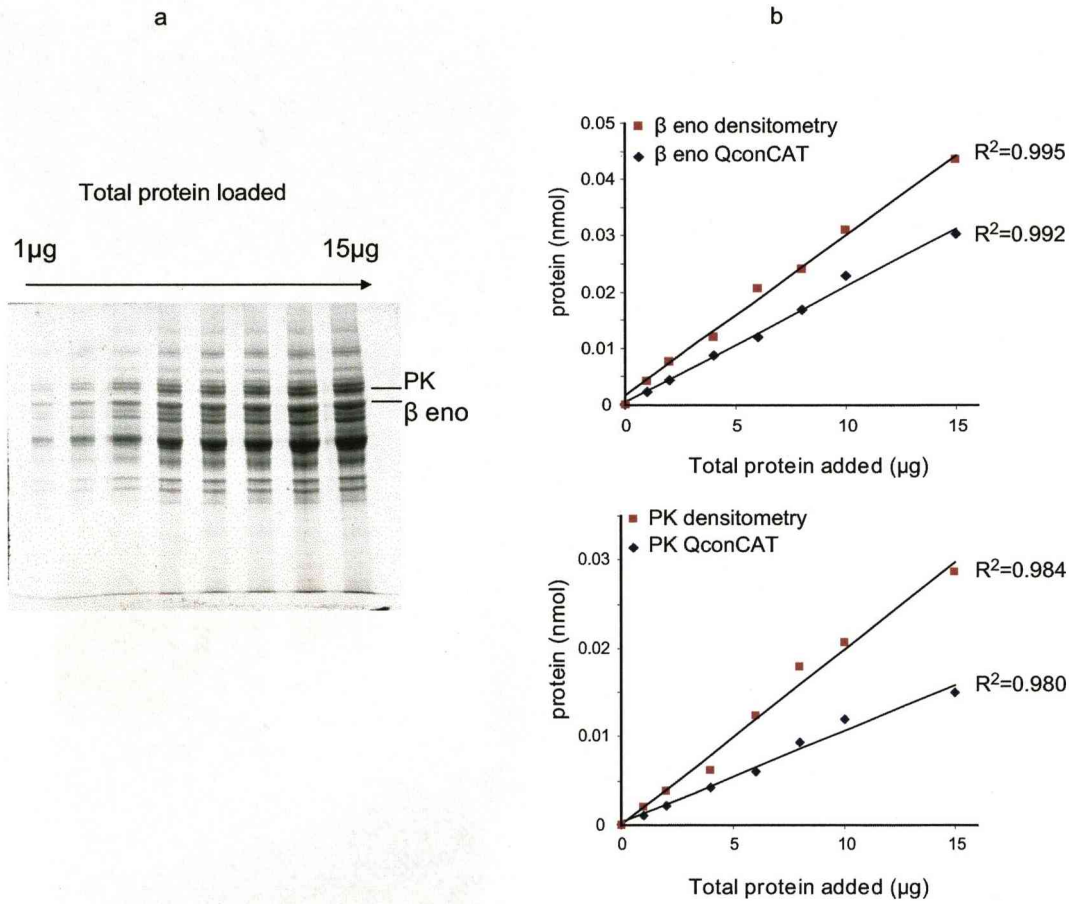



**Figure 66. Quantification of increasing amount of total protein, by densitometry and the QconCAT method.**

Chicken skeletal muscle soluble protein was analysed by SDS-PAGE with increasing amounts of total protein (µg) loaded onto the gel (a). The amounts of two proteins, beta enolase (β eno) and pyruvate kinase (PK; nmol/g tissue) were measured using both QconCAT and densitometry. For each protein quantified using both methods, the correlation coefficient ($R^2$) is expressed to the right of the graph.

To assess the extent to which intact mass analysis by ESI-Q-ToF MS reflects relative abundance of proteins in complex mixtures, protein quantification using QconCAT was compared to relative signal intensity of chicken skeletal muscle soluble proteins from a 30d broiler when analysed by ESI-Q-ToF MS (data acquired by Dr. J. Hayter). For seven of the most abundant proteins identified in the deconvoluted mass spectrum, protein abundance was expressed as a percentage signal intensity of the total intensity of the seven proteins in the processed mass spectrum. To compare with absolute quantification by QconCAT, the absolute amount of each protein was expressed as a percentage of the total for the seven proteins (Figure 67). Correlation between the two techniques was good ($R^2$=0.9264), highlighting the potential value of intact mass analysis to give an indication of protein abundance. Relative abundance is over-estimated from intact mass analysis by ESI-Q-ToF MS, compared to QconCAT, but this is difficult to interpret from the results of a single comparison. If this were a consistent observation from several comparisons, of samples varying in complexity and abundance of each protein, it may be taken into account for more reliable quantification. However, as a measure of relative quantification, this is not necessary, thus intact mass analysis provides a more reliable method than densitometry analysis following 1D SDS-PAGE. When comparing quantification by all three methods of intact mass analysis by ESI-Q-ToF MS, densitometry from 1D SDS-PAGE and QconCAT (Figure 68), it is clear that densitometry does not compare well with intact mass ($R^2$=0.4813) or QconCAT ($R^2$=0.5848) and is consequently not the method of choice for reliable measurements of relative protein abundance. This is most probably due to the different affinity of individual proteins for the stain in addition to co-migration of multiple proteins through the gel.

## 5.5 Absolute Quantification of Chicken Skeletal Muscle Soluble Proteins

The QconCAT strategy for absolute quantification of known skeletal muscle proteins was achieved by adding a known amount of [$^{13}C_6$]arg/[$^{13}C_6$]lys-labelled QconCAT protein to chicken skeletal muscle soluble proteins, digesting with trypsin and comparing signal intensities of chicken skeletal muscle peptides with corresponding labelled internal standard peptides in mass spectra. QconCAT protein was added in a 1:10 (QconCAT:chicken skeletal muscle protein) ratio to chicken skeletal muscle soluble fraction samples taken from both broiler and layer strains at six time points during growth. For each time point, four birds were analysed. This ratio was selected pragmatically based on the abundance of the major soluble proteins in

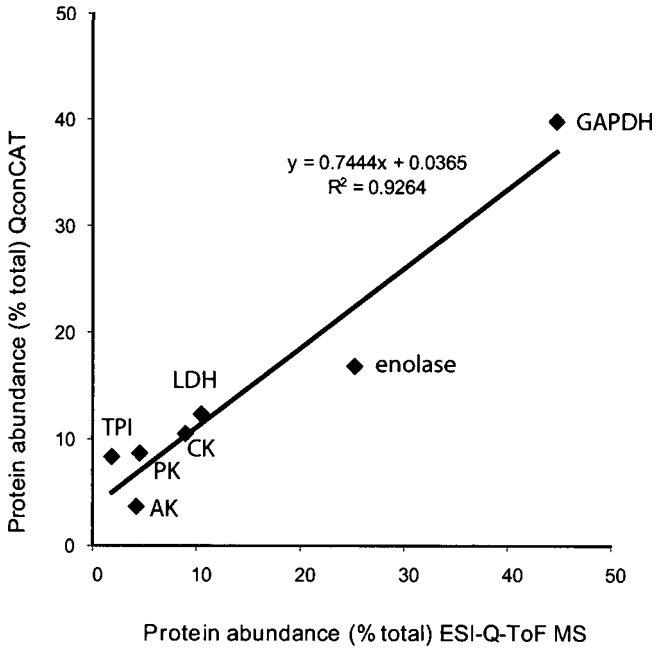



**Figure 67. Comparison of quantification using intact mass analysis by ESI-Q-ToF MS and the QconCAT method.**

Chicken skeletal muscle soluble proteins from a 30d broiler were analysed by ESI-Q-ToF MS. As a measure of abundance, the relative signal intensity of each protein peak (when processed to give the true mass of each protein using MaxEnt1; MassLynx) was expressed as a percentage of the total intensity of all seven proteins in the processed mass spectrum (experimental data obtained by Dr. J. Hayter). To compare with quantification by QconCAT, the absolute amount of each protein was expressed as a percentage of the total for the seven proteins.

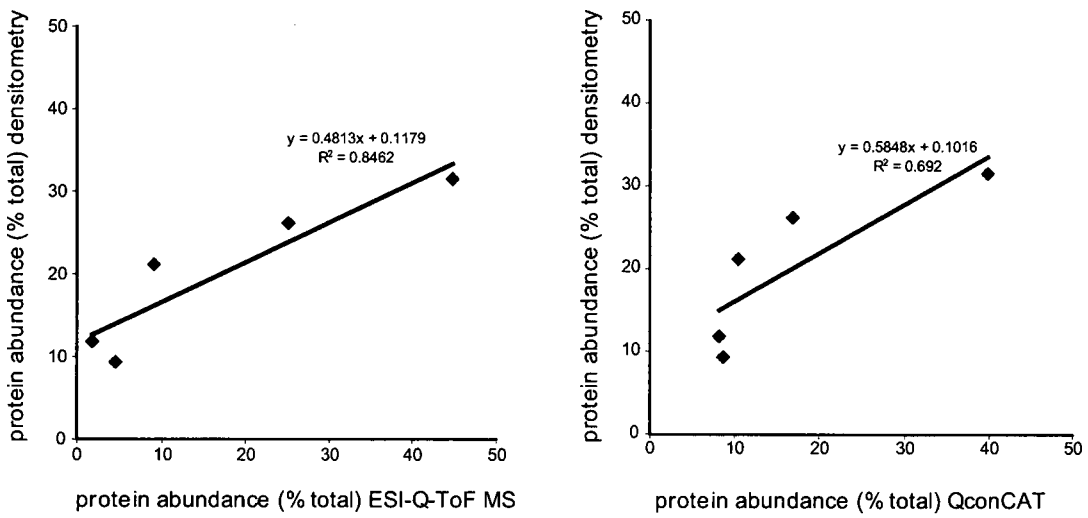



**Figure 68. Comparison of quantification, using intact mass analysis by ESI-Q-ToF MS, SDS-PAGE and densitometry analysis, and the QconCAT method.**
Chicken skeletal muscle soluble proteins from a 30d broiler were analysed by ESI-Q-ToF MS. As a measure of abundance, the relative signal intensity of each protein peak (when processed to give the true mass of each protein using MaxEnt1; MassLynx) was expressed as a percentage of the total intensity for the processed mass spectrum (experimental data obtained by Dr. J. Hayter). To compare with quantification by QconCAT and densitometry analysis of 1D SDS-PAGE (experimental data obtained by I.J. Edwards), the absolute amount (QconCAT) and relative amount (densitometry) of each protein were expressed as percentages of the total for the seven proteins.

chicken skeletal muscle in all muscle samples from 1-30d growth (Figure 69). After co-digestion of chicken skeletal muscle soluble proteins and [$^{13}C_6$]arg/[$^{13}C_6$]lys-labelled QconCAT with trypsin, MALDI-ToF MS analysis of peptides produced highly complex mass spectra. However, 10 out of 20 Q-peptides could be identified in the composite spectrum without further sample processing and were therefore used for quantification. For these 10 proteins, for example glyceraldehyde 3-phosphate dehydrogenase (GAPDH; Figure 70), the change in protein expression was measured during growth from 1d to 30d post hatch by converting relative signal intensities of analyte and internal standard peptide ions into absolute quantities of analyte protein, expressed as nmol/g net weight breast muscle tissue. For GAPDH, the dramatic increase in protein expression is more apparent when the spectra are normalised to a constant intensity of the internal standard. This is compared with 1D SDS-PAGE analysis in which a constant 10µg total protein was applied to each lane from birds 1-30d post-hatch. For the identification and quantification of further proteins, not quantifiable by MALDI-ToF MS, ionisation of lysine terminated peptides was improved by guanidination, permitting quantification of embryonic myosin (E myo) and triose phosphate isomerase (TPI). Furthermore, peptide mixtures (chicken skeletal muscle soluble proteins and QconCAT) were analysed by LC-ESI-Q-ToF MS, permitting quantification of five proteins previously quantified by MALDI-ToF MS (with and without guanidination), in addition to six proteins previously not detected. For absolute quantification of chicken skeletal muscle soluble proteins using the QconCAT method for the complete biological study, all proteins quantified by MALDI-ToF MS (with and without guanidination) and LC-ESI-Q-ToF MS were expressed as nmol/g of pectoralis muscle tissue. Data were obtained during growth from 1 to 30 days post-hatch for four birds at each time point for chickens of the layer and broiler strains (Figure 71). For proteins quantified using multiple methods of analysis, data have been plotted separately, adjacent to the alternative data set and the similarity between the two is clearly represented. Spectral data is also included for proteins adenylate kinase (AK), quantified by MALDI-ToF MS, alpha actin, quantified by LC-ESI-Q-ToF MS and tropomyosin A (TM A), quantified as 0nmol/g tissue by LC-ESI-Q-ToF MS. For the biological study, changes in protein expression for 17 proteins incorporated into the QconCAT protein for absolute quantification are displayed for both broiler and layer chickens. This permitted a simple comparison of individual proteins, highlighting that some demonstrate massive pool expansion, whilst others declined to a similar degree. Within the realms of this single QconCAT experiment, a measurable dynamic range across all proteins was covered, of 10nmol/g to 550nmol/g for a single protein (GAPDH) and as low as 2±1nmol/g (α eno; 1d broiler). Thus, protein concentrations over a 300-fold range were assessed. This

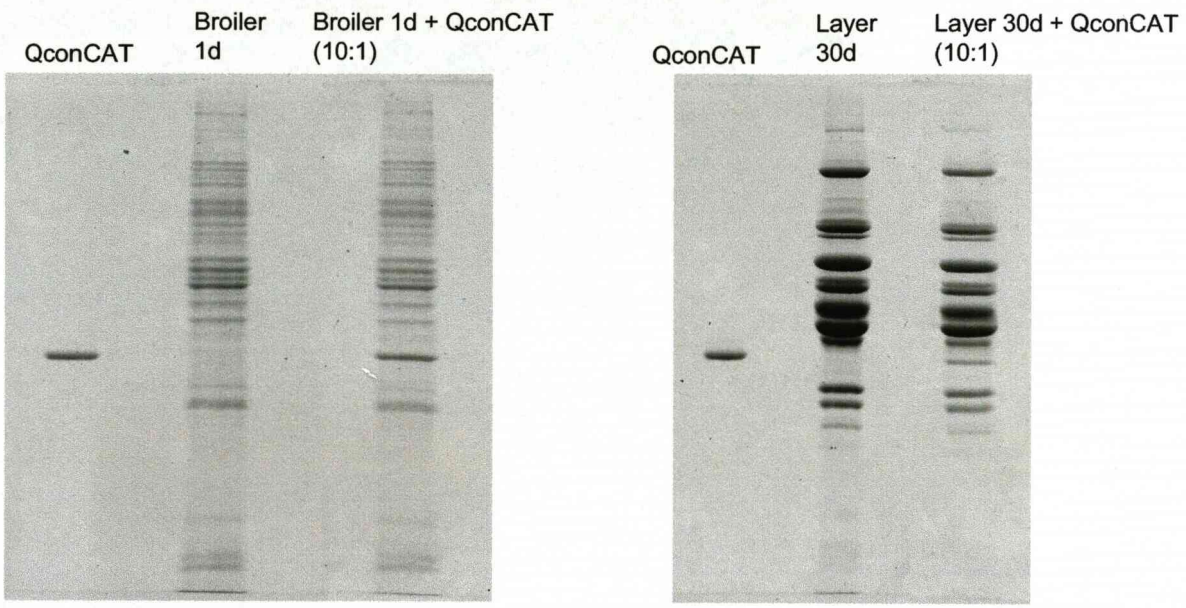



**Figure 69. SDS-PAGE analysis of QconCAT and chicken skeletal muscle soluble proteins from broiler and layer strains.**

1d broiler and 30d layer chicken skeletal muscle was homogenised and the soluble proteins were analysed by SDS-PAGE. QconCAT protein was added to the skeletal muscle preparation in a 1:10 ratio by total protein concentration and the mixture was analysed by SDS-PAGE.

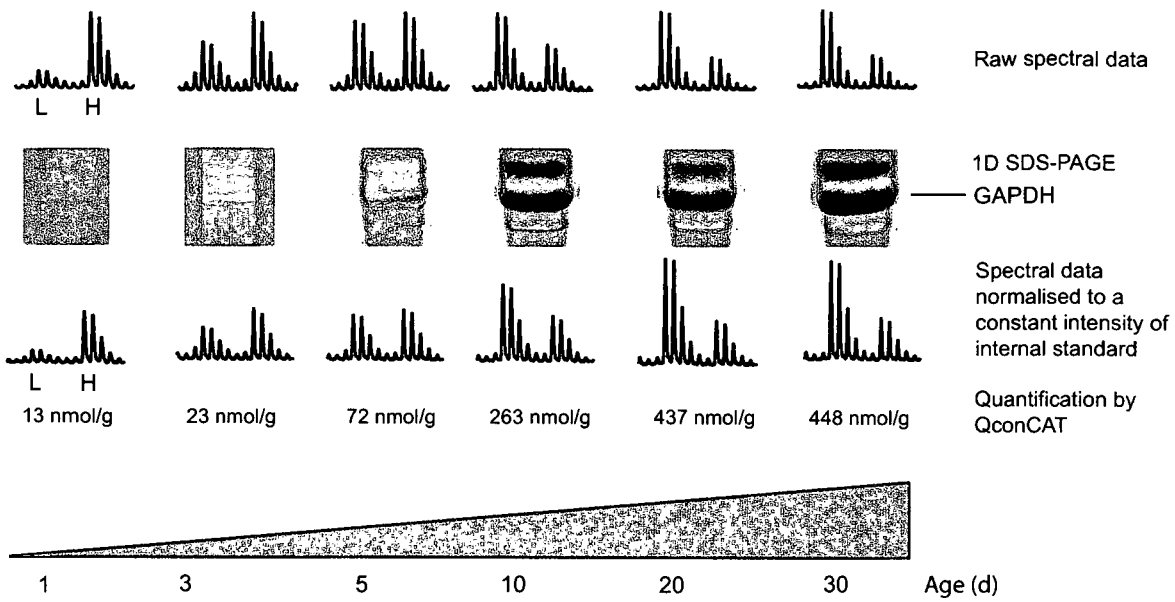



**Figure 70. Quantification of GAPDH expression in chicken skeletal muscle.**
Soluble muscle proteins were prepared from pectoralis skeletal muscle of birds from 1d to 30d post hatching. Each sample (70µg of protein) was mixed with a constant amount of QconCAT (7µg) and digested to completion with trypsin before analysis by MALDI-ToF MS. The change in expression is measured using the relative peak intensity of the analyte and internal standard peptide at each time point. The dramatic increase in protein expression is more apparent when the spectra are normalised to a constant intensity of the internal standard. This change in protein expression is also apparent by 1D SDS-PAGE analysis of chicken skeletal muscle soluble proteins, in which a constant 10µg of total protein was applied to each lane. The amount of GAPDH at each time point during growth is also expressed as nmol/g tissue as quantified using QconCAT.

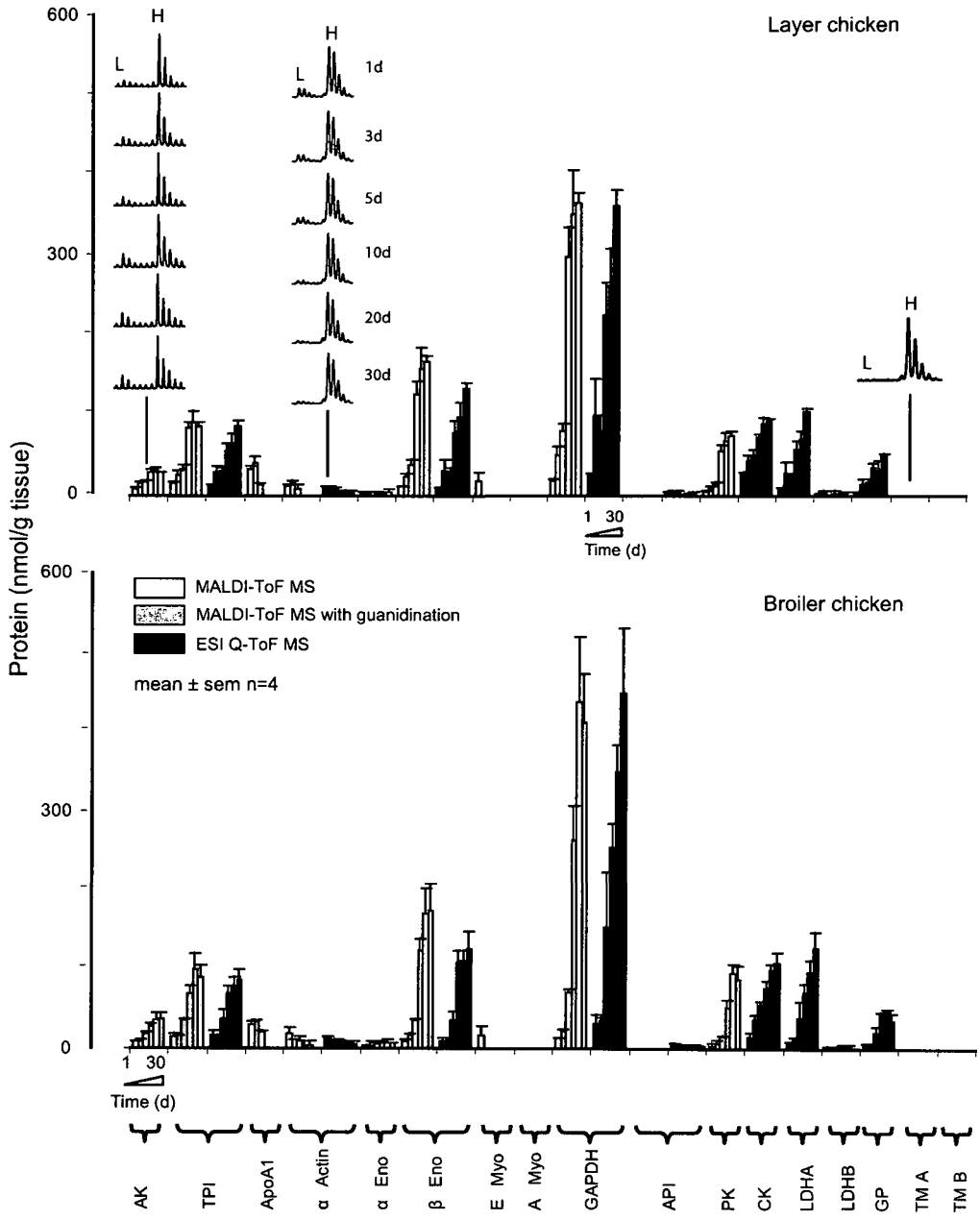



**Figure 71. Quantification of chicken skeletal muscle protein expression by QconCAT.** Soluble protein derived from broiler and layer chickens (70µg) was mixed with QconCAT protein (7µg) and digested to completion with trypsin. The digests were analysed by MALDI-ToF MS (with or without guanidination) or LC-ESI-Q-ToF MS. For five proteins (triose phosphate isomerase; TPI, alpha actin, beta enolase; β eno, glyceraldehyde 3-phosphate dehydrogenase; GAPDH and actin polymerisation inhibitor; API) multiple methods were used to quantify a single protein during growth; these data have been plotted separately, adjacent to the alternative data set and these have been grouped below the x-axis. Each cluster of data represents six time points during growth (1d, 3d, 5d, 10d, 20d and 30d) for four birds of each strain at each time point. The data are presented as the absolute tissue amount (nmol/g tissue) and expressed as mean±sem. Mass spectra are included for proteins adenylate kinase (AK), alpha actin and tropomyosin A (TM A) to highlight the difference in relative signal intensity. For proteins expressed as 0nmol/g, ions corresponding to analyte peptides were not present in the spectrum (see spectral data for TM A).

quantification can be subtle, for example in monitoring isoform changes from embryonic to adult myosin, as well as a change in state from free, soluble protein to that assembled within the myofibrillar apparatus (actin). It is also possible to monitor expression of isoforms of the same enzyme for which Q-peptides differ only in a single amino acid (lactate dehydrogenases A and B). These data were also used for a comparison of protein expression in both broiler and layer strains, which demonstrate a dramatic difference in growth rate of the pectoralis muscle (section 1.5.1). However, comparing absolute amounts of abundant proteins between broiler and layer strains (Figure 72), showed very little or no difference in protein expression during growth per gram of muscle tissue. The increased weight gain observed in broiler chickens must therefore be attributable to total pool expansion of the breast muscle, rather than an increase in the amount of individual proteins relative to the layer.

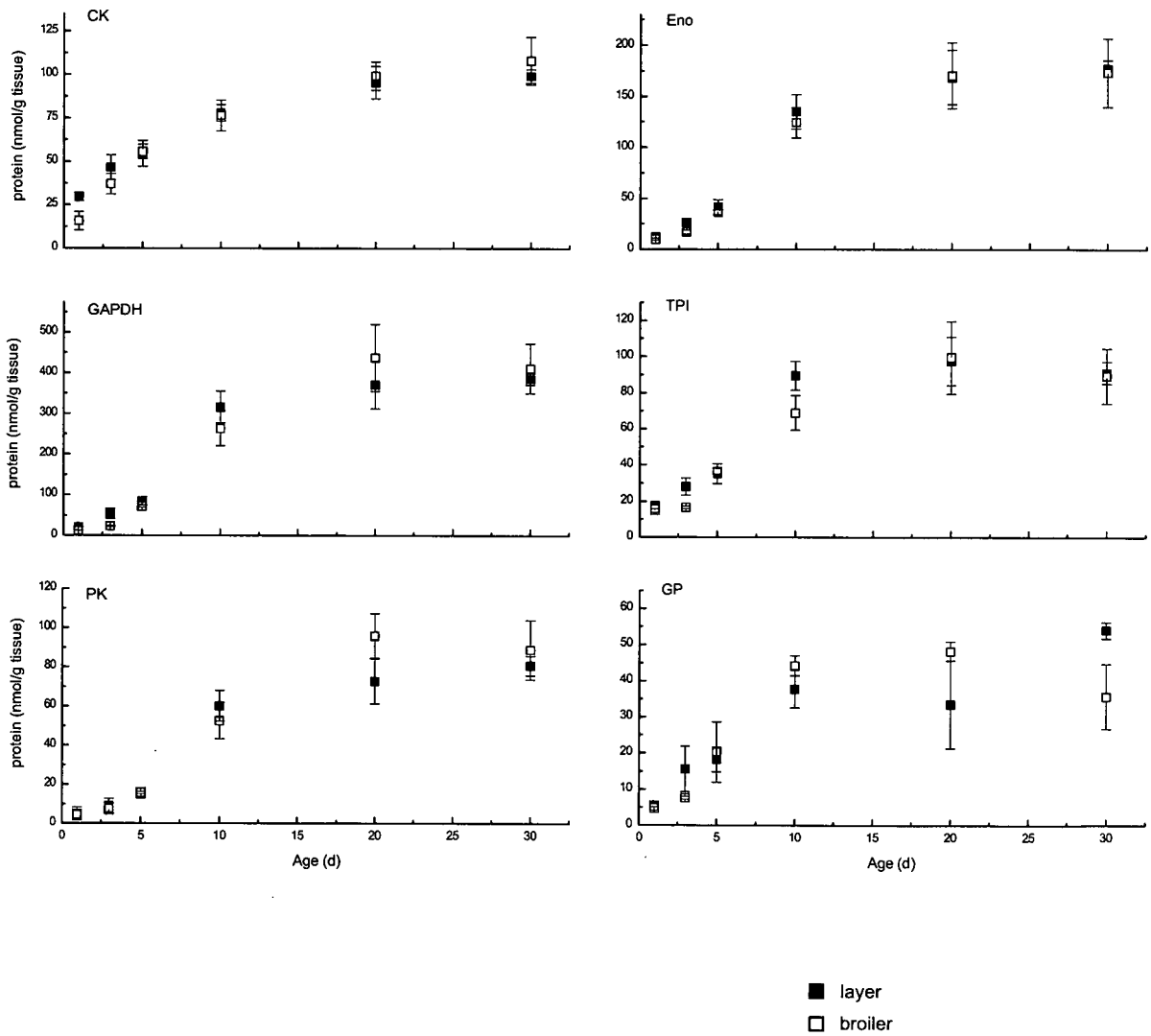



**Figure 72. Comparison of absolute protein quantification of soluble skeletal muscle proteins in broiler and layer chickens.**

Absolute quantification of chicken skeletal muscle soluble proteins was achieved using the QconCAT method. For six proteins, data for broiler (open squares) and layer (closed squares) strains were compared. Each data point is plotted as mean±sem n=4.

## 5.6 Additional Applications of QconCAT Technology

### 5.6.1 Quantification of soluble skeletal muscle proteins in other species

To investigate the application of the QconCAT method for quantification of soluble muscle proteins in other species, proteins glyceraldehyde 3-phosphate dehydrogenase (GAPDH), pyruvate kinase (PK) and adenylate kinase (AK) purified from rabbit skeletal muscle were quantified in known amounts using the QconCAT protein designed for chicken skeletal muscle (Figure 73). For GAPDH and PK, sequences of peptides represented in the QconCAT designed for chicken skeletal muscle differed by a single amino acid, and AK had the same peptide sequence and was used as a control. From 1D SDS-PAGE analysis, AK purified from rabbit also contained a considerable amount of creatine kinase (CK; supplementary figure 23); the relative proportion of these two proteins was assessed by densitometry to correct for the amount of AK added (pmol). For all three proteins, correlation was strong; AK $R^2$=0.998 slope=0.995, PK $R^2$=0.994 slope=0.981, GAPDH $R^2$=0.990 slope=1.19. The single amino acid substitution for GAPDH and PK did not make a significant difference to the quantification but this would depend on the amino acid that is substituted, for example if an arginine was substituted for a lysine, this would have a much greater effect on the signal intensity relative to the standard peptide, as previously discussed (section 3.3.2). To quantify proteins from other species in context, soluble skeletal muscle proteins from carp, mouse and chicken were analysed by 1D SDS-PAGE (Figure 74). It is clear from this analysis that several skeletal muscle proteins are highly conserved among these species and proteins GAPDH and CK were identified in carp and mouse samples by in-gel digestion with trypsin and peptide analysis by MALDI-ToF MS. Peptide mass fingerprinting (supplementary figures 24&25) highlighted sequence differences between species for these proteins, particularly for Q-peptides used for quantification of chicken skeletal muscle soluble proteins. For quantification, carp and mouse skeletal muscle soluble proteins were independently mixed with a known amount of QconCAT and co-digested in-solution with trypsin (protein:trypsin 20:1). Peptides were analysed by LC-ESI-Q-ToF MS with reversed-phase peptide separation over a 50min acetonitrile gradient (0-100%). Ion chromatograms for GAPDH and CK from both species (carp and mouse) were extracted and combined to give mass spectra (Figure 75). The signal intensities of analyte and standard peptides were used for quantification. Quantification for these proteins achieved using QconCAT was compared to that using densitometry. Given that agreement between these two methods for quantification is poor (as previously discussed; section 3.4.4) it is difficult to draw reliable conclusions from this cross-species analysis. The two methods for quantification of CK and GAPDH in carp and mouse skeletal muscle gave different results, suggesting that although

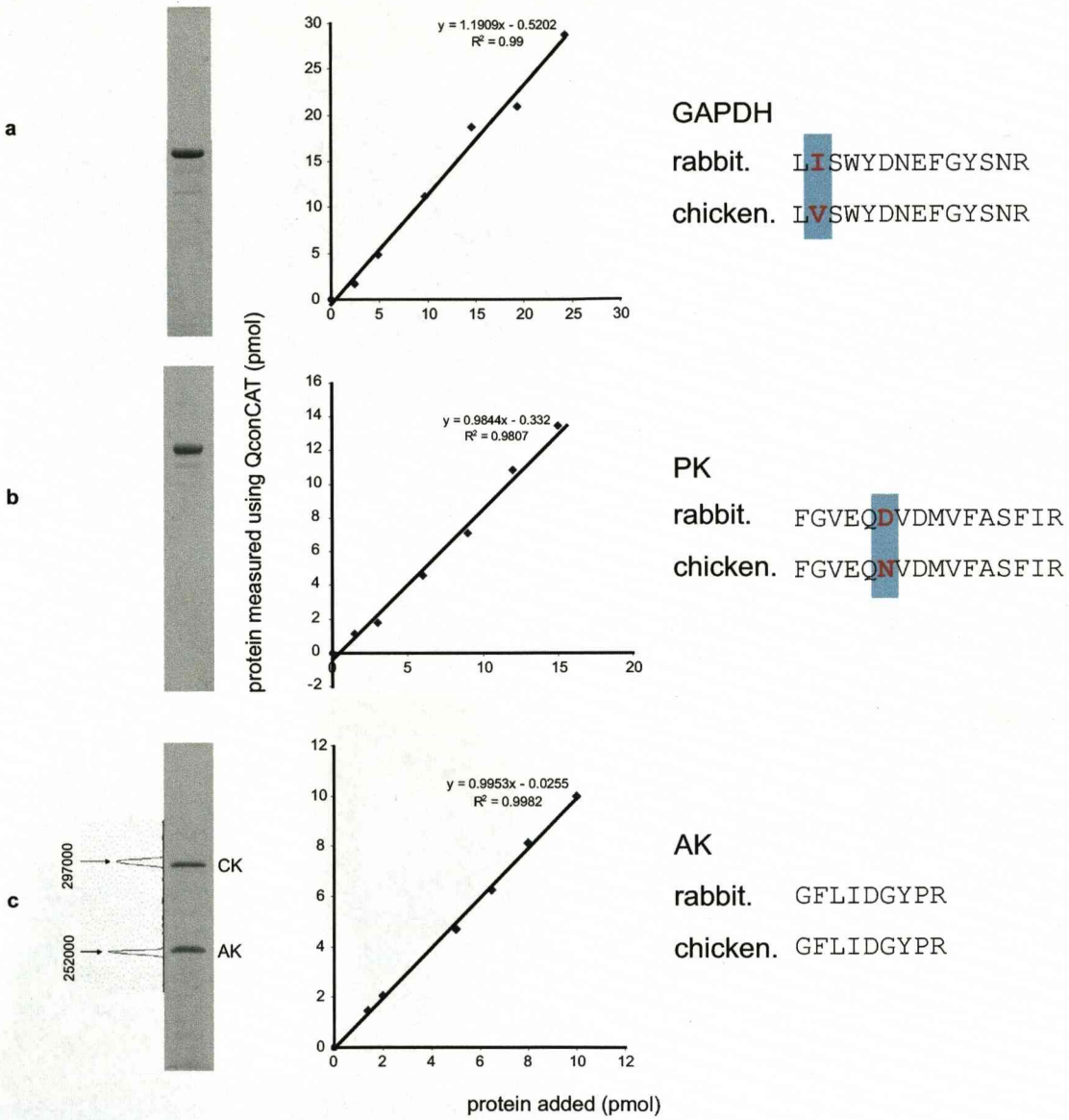



**Figure 73. Quantification of skeletal muscle proteins from rabbit and chicken by QconCAT.** Purified proteins glyceraldehdye 3-phosphate dehydrogenase (GAPDH), pyruvate kinase (PK) and adenylate kinase (AK) from rabbit were analysed by 1D SDS-PAGE, mixed with QconCAT and digested in-solution with trypsin. Purified proteins were added from 0-10pmol (AK), 0-15pmol (PK), 0-25pmol (GAPDH) to QconCAT protein from 10 to 100pmol. Peptides were analysed by MALDI-ToF MS and the relative signal intensity of analyte and internal standard ions was used for quantification. For each protein, GAPDH (a), PK (b) and AK (c), the sequence of the Q-peptide in both rabbit and chicken are indicated to the right with amino acid differences highlighted. For AK, the 1D SDS-PAGE image was analysed by densitometry to assess the proportion of the two main bands. These were identified as unique proteins; creatine kinase (CK) and adenylate kinase (AK) by peptide mass fingerprinting and quantification was corrected accordingly (for mass spectra and peptide mass fingerprinting see supplementary figure 23).

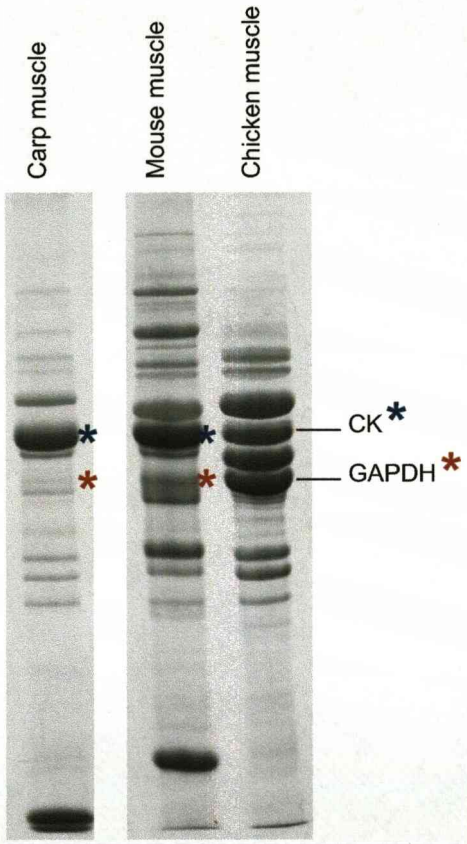



**Figure 74. Skeletal muscle soluble proteins from carp, mouse and chicken.**

Soluble skeletal muscle proteins of mouse and carp (3mg/mL, prepared by L. McDonald and L. McLean) were analysed by 1D SDS-PAGE. Proteins creatine kinase (CK) and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) were identified by in-gel digestion with trypsin and peptide mass fingerprinting by MALDI-ToF MS (supplementary figures 24&25). To illustrate corresponding proteins, chicken skeletal muscle soluble proteins were analysed simultaneously and CK is indicated by a blue star to the right in each species, and GAPDH a red star.

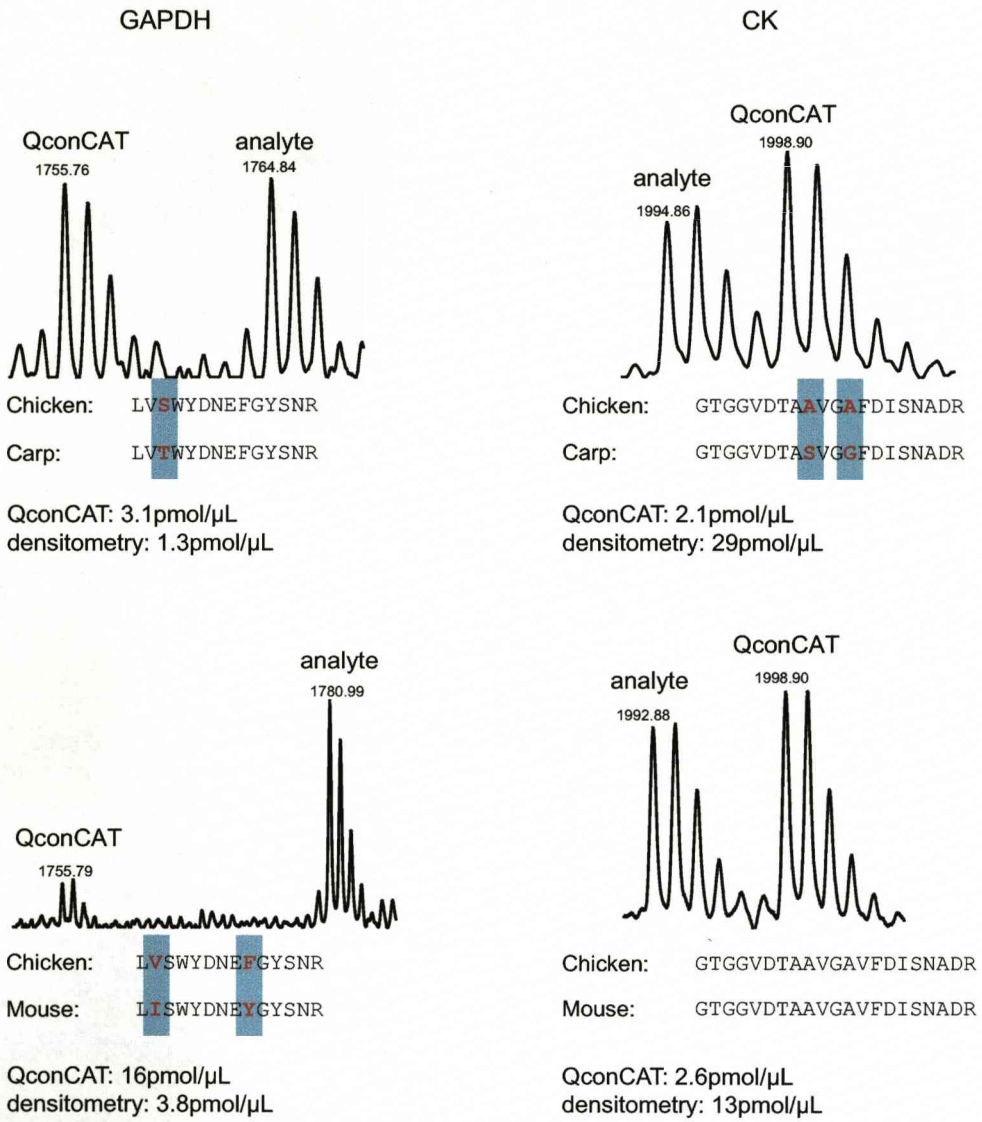



**Figure 75. Analyte: internal standard peptide pairs for cross-species quantification.**
Carp and mouse skeletal muscle soluble proteins (15µg, prepared by L. McDonald and L. McLean) were co-digested with QconCAT protein (5µg) in solution with trypsin at a ratio of protein: enzyme of 20:1. Peptides were analysed by LC-ESI-Q-ToF MS and Q-peptide ions from glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and creatine kinase (CK) with 0, 1 or 2 amino acid substitutions (highlighted in peptide sequences underneath mass spectra) were used for quantification relative to the signal intensity of the corresponding QconCAT peptide ion.

amino acid substitutions valine for isoleucine and asparagine for aspartic acid, in purified proteins GAPDH and PK had no significant effect on ionisation and subsequent quantification by QconCAT (Figure 73), the substitution serine for threonine, or two substitutions of alanine for serine and glycine (carp CK) and valine and phenylalanine for isoleucine and tyrosine (mouse GAPDH) had a significant effect on quantification. Alternative explanations could include problems previously discussed for 1D SDS-PAGE and densitometry, for example co-migration of several proteins to the same location on the gel or incomplete digestion of analyte proteins as this was not investigated in this context. Additionally there may be a biological explanation, for example proteolytic processing of mature proteins which is documented in carp skeletal muscle when animals are acclimated to different temperatures (McLean *et al.,* 2007), or stable post-translational modifications impacting on quantification. For a typical QconCAT experiment, a comprehensive analysis of analyte proteins would be conducted prior to design and implementation of such a strategy, thus ensuring reliable quantitative data. Assessment of accuracy in quantification of muscle proteins from other species is difficult, thus conclusions that can be drawn from this experiment are limited. However, it is clear that quantification can be achieved using QconCAT where the analyte peptide has one amino acid difference as achieved for purified proteins from rabbit skeletal muscle. In context there was no way of accurately reconciling the quantitative data acquired with the amount of each protein in the tissue, thus the extent to which this remained reliable was unknown. This will depend on the amino acid difference and how likely this is to affect signal intensity. For many proteins, especially those that have highly conserved peptide sequences that only differ by one or two amino acids, the effects on quantification could be rigorously tested using a number of purified proteins or synthetic peptides containing the specific sequence difference. This could be used to adjust signal intensity data, achieving absolute quantification of multiple species using the same QconCAT protein. Alternatively this could be built into the design criteria, to select where possible peptides that are identical in several species, for example the N-terminal peptide which is often the most highly conserved.

### 5.6.2 Quantification of normalisation using Equalizer™ bead technology

In an approach to reduce dynamic range of complex protein samples, Equalizer™ bead technology was developed whereby high abundance proteins are diminished and low abundance or trace proteins are enriched simultaneously by binding to a library of ligands exposed on the surface of beads (Thulasiraman *et al.,* 2005, Righetti *et al.,* 2006). This application, whilst not relevant for quantitative protein analysis, has been tested using the

soluble fraction of chicken skeletal muscle, with use of the QconCAT protein an ideal opportunity to quantify normalisation. For normalisation, 20mg Prospectrum-2 beads were washed and swollen in 50% (v/v) methanol, prior to equilibration in 20mM sodium phosphate buffer. Chicken skeletal muscle soluble proteins, initially 25mg were exposed to the beads for 2h, after which beads were washed extensively to remove unbound protein. For efficient normalisation, analyte proteins must saturate all available ligand binding sites on the surface of the bead library (Guerrier *et al.*, 2006), thus 25mg, 50mg and 100mg total protein were each incubated with 20mg beads as described above. Beads were washed to remove unbound protein prior to 1D SDS-PAGE analysis of starting material (chicken skeletal muscle soluble proteins), beads and unbound protein (Figure 76). Wash fractions containing unbound protein were analysed in the same way to ensure that these had been removed (results not shown). From 1D SDS-PAGE analysis, protein normalisation was improved with an increased amount of protein exposed to the beads, reflecting efficient saturation of ligands carried on the beads, resulting in relative enrichment of low abundance proteins, bringing them to the range where they can be visualised on a 1D gel. Simultaneously, the highly abundant proteins are progressively suppressed and do not dominate the preparation at higher loadings, although at lower loadings, they remain abundant. It is apparent that the amount of each protein bound to the beads varies, even when 100mg protein are incubated with 20mg beads, suggesting that some proteins may have a greater affinity for the beads. In fact, the most abundant band on the gel of normalised material after 100mg protein was exposed to the beads was not observed before normalisation. 1D SDS-PAGE analysis of beads exposed to increasing amounts of protein was also used for densitometry analysis of the major gel bands, highlighting the level of normalisation compared to the starting material as the dominance of a few abundant proteins in the starting material is removed giving rise to many more bands of more equal stain intensity (Figure 77). To identify these proteins, particularly those that have changed as a result of normalisation, gel lanes containing starting and normalised material were cut into 22 slices and each gel slice was digested with trypsin. Peptides were analysed by LC-ESI-LTQ MSMS from which peptide sequences were searched against MSDB using the MASCOT search engine. Proteins identified with a MOWSE score of greater than 45 were accepted as confident identifications ($p<0.05$; Figure 78, Table 5). After normalisation, many more proteins were identified, in particular, the most intense band on the gel in the normalised material (after exposure to 100mg total protein) was glucose-6-phosphate isomerase, a protein that was not identified without normalisation and thus was greatly enriched following exposure to the beads. Although observed in a single experiment, this selective enrichment was also clear upon

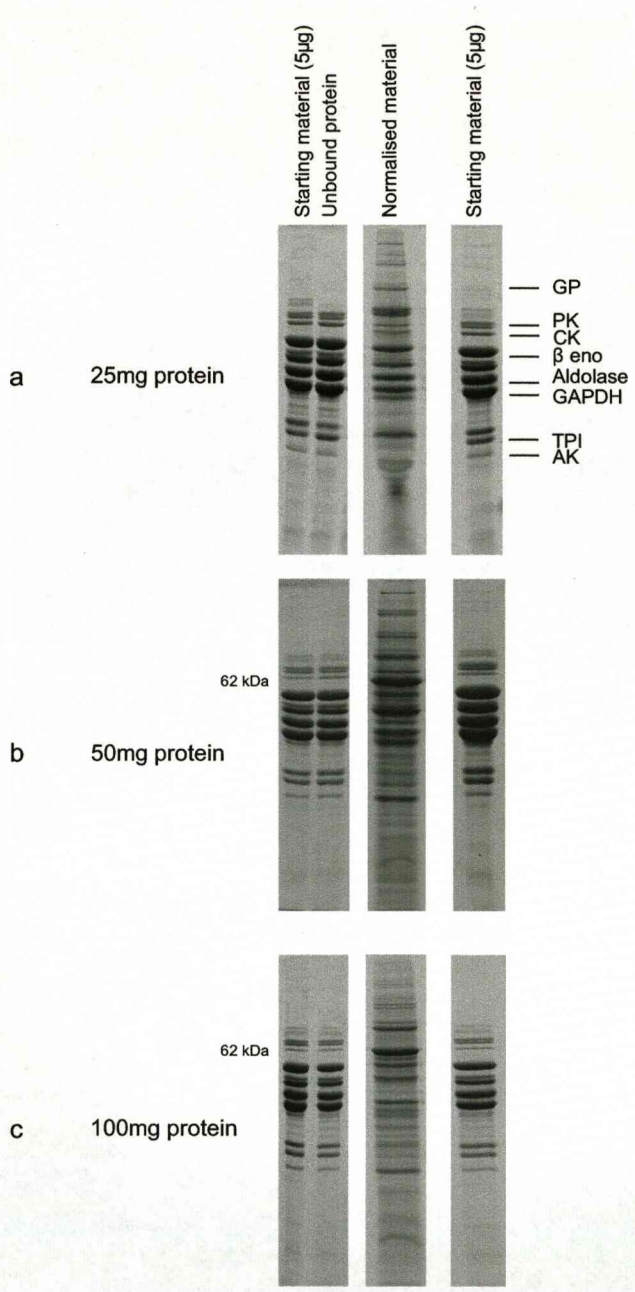



**Figure 76. Normalisation of chicken skeletal muscle soluble protein abundance using Equalizer ™ beads.**

To isolate the soluble fraction of chicken skeletal muscle, supermarket purchased chicken breast tissue (2g) was homogenised in 18mL 20mM sodium phospha te buffer, pH7.0 containing protease inhibitors (Complete Protease Inhibitors, Roche, Lewes, UK). This was centrifuged at 15,000 x *g* for 45 minutes at 4°C. The supernatant fraction, containing soluble protein, was then removed. The total protein concentration of the final preparation was measured using a Coomassie Plus Protein Assay (Pierce, Northumberland, UK). For normalisation, 20mg Prospectrum-2 (Louisville, KY, USA) beads were washed in 1mL 50% (v/v) MeOH and mixed gently for 10min. Beads were allowed to settle and the supernatant was removed and discarded. MeOH 50% (v/v) was added to cover the surface of the beads that were left to swell overnight at 4°C. Once swollen, 20mg beads (constituting 100µL settled bed volume) were transferred to a 1.5mL Eppendorf tube. Beads were washed in 1mL double distilled $H_2O$ on a roller mixer fo 30min prior to equilibration by repeated washing in 20mM sodium phosphate buffer pH7.0 for 30min. After each wash, beads were left to settle for 5min and the supernatant was removed. Approximately 1mL sample containing 25mg (a), 50mg (b) and 100mg (c) soluble protein in three separate experiments was added to the beads and mixed for 2h on a roller mixer. Unbound protein was collected as the supernatant fraction after beads had settled for 5min. The beads were subsequently washed eight times in 1mL 20mM phosphate buffer and supernatant fractions were removed and collected. Starting material, unbound protein, wash fractions and beads containing bound, normalised protein were analysed using 1D SDS-PAGE. Equalizer™ beads were loaded and run directly with no prior protein elution.

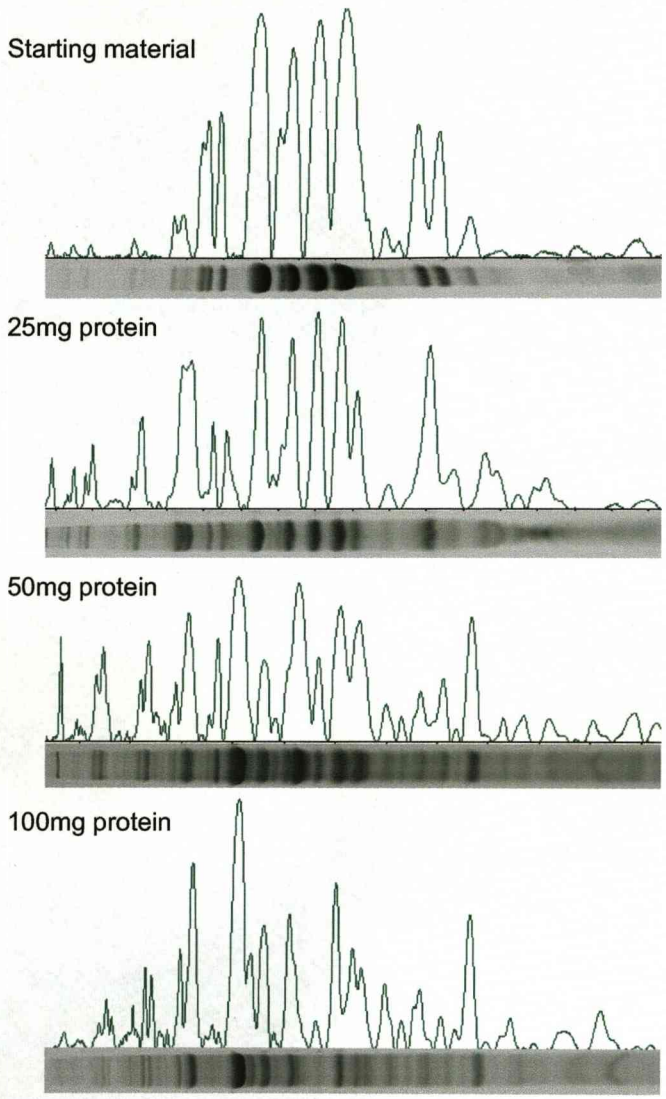



**Figure 77. Densitometry analysis of 1D SDS-PAGE separated chicken skeletal muscle soluble proteins normalised using Equalizer™ beads.**
Densitometry analysis was applied to 1D SDS-PAGE images from starting material and normalised material when 25mg, 50mg and 100mg chicken skeletal muscle soluble proteins were exposed to 20mg Equalizer™ beads. For each analysis, the densitometry trace measuring band volume is aligned with the gel image below.

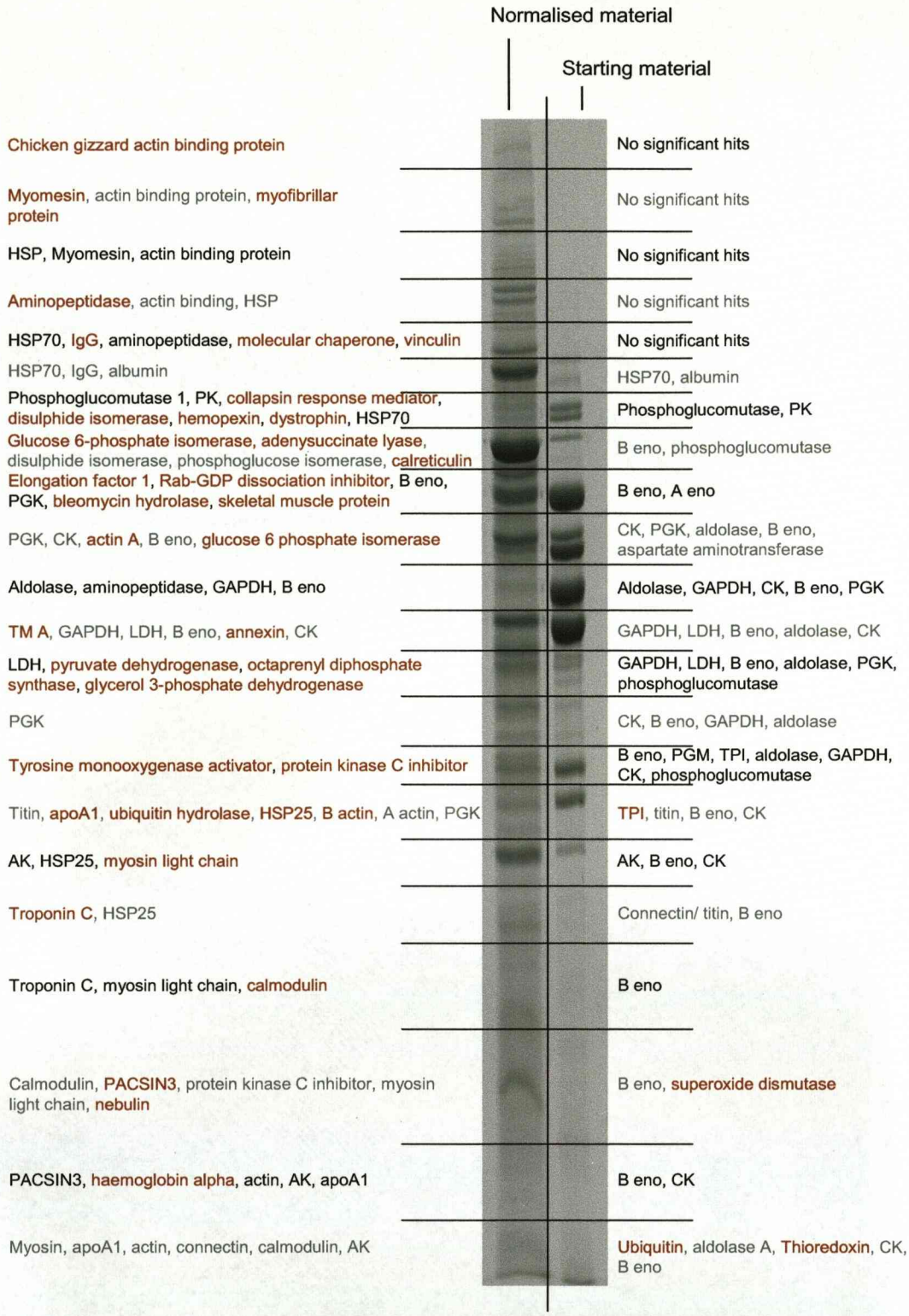



**Figure 78. Identification of normalised proteins using in-gel digestion with trypsin and LC-ESI-LTQ MSMS.** For protein identification, 1D SDS-PAGE separations of starting material and beads containing normalised proteins were divided into 22 slices, each of which was de-stained using 50:50 acetonitrile:50mM ammonium bicarbonate, dehydrated with acetonitrile and digested overnight in-gel with trypsin. Resulting peptide solutions were diluted 1:50 with 0.1% (v/v) formic acid and analysed by LC-ESI-LTQ MSMS using an Ultimate 3000 HPLC system (Dionex, UK) coupled to a LTQ (Thermo Finnigan, UK). Nanoflow HPLC at 300nL/min was used to resolve peptides (in 0.1% v/v formic acid) over a 60 minute acetonitrile gradient (0-100%). Peptides were acquired over the mass range 400-1500m/z with the capillary voltage set at 50V, spray voltage at 1.8kV. MSMS data were searched against MSDB using MASCOT from which only confident identifications (MOWSE score>45, $p<0.05$) were accepted, for details see table 5 (overleaf).

| | Chicken skeletal muscle soluble proteins | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gel band # | Protein | MOWSE score | Swissprot accession # | Molecular weight (Da) | First species | # peptides matched | % coverage | Sequence of single peptide match |
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | Heat shock protein 70 | 162 | Q3MHM4 | 71226 | *Bos taurus* | 8 | 12 | |
| | Albumin | 47 | P19121 | 69872 | *Gallus gallus* | 5 | 7 | |
| 7 | Phosphoglucomutase | 632 | Q2UZR2 | 66550 | *Gallus gallus* | 35 | 39 | |
| | Pyruvate kinase | 248 | P00548 | 57978 | *Gallus gallus* | 22 | 30 | |
| 8 | Glucose 6-phosphate isomerase | 105 | Q5ZMU3 | 62177 | *Gallus gallus* | 11 | 13 | |
| | Phosphoglucomutase | 94 | Q2ZUZR2 | 66550 | *Gallus gallus* | 6 | 12 | |
| | Beta enolase | 69 | P13929 | 46826 | *Homo sapiens* | 2 | 3 | |
| 9 | Beta enolase | 1043 | P07322 | 46839 | *Gallus gallus* | 52 | 52 | |
| | Creatine kinase | 55 | P00565 | 43301 | *Gallus gallus* | 1 | 3 | K.LSVEALNSLEGEFK.G |
| | Alpha enolase | 47 | Q9W6DO | 40415 | *Sphenodon punctatus* | 7 | 5 | |
| 10 | Creatine kinase | 1022 | P00565 | 43301 | *Gallus gallus* | 57 | 39 | |
| | Phosphoglycerate kinase | 689 | P51903 | 44557 | *Gallus gallus* | 32 | 34 | |
| | Beta enolase | 108 | P07322 | 46839 | *Gallus gallus* | 2 | 8 | |
| | Aspartate aminotransferase | 95 | P00504 | 45775 | *Gallus gallus* | 2 | 7 | |
| | Aldolase | 91 | Q76BC5 | 35908 | *Cephaloscyllium umbratile* | 1 | 3 | R.LAIMENANVLAR.Y |
| 11 | Aldolase | 443 | Q76BEO | 35952 | *Acipenser baerii* | 15 | 15 | |
| | Creatine kinase | 211 | P00565 | 43301 | *Gallus gallus* | 9 | 30 | |
| | Glyceraldehyde 3-phosphate dehydrogenase | 81 | P00356 | 35550 | *Gallus gallus* | 4 | 19 | |
| | Beta enolase | 59 | P07322 | 46839 | *Gallus gallus* | 2 | 10 | |
| 12 | Glyceraldehyde 3-phosphate dehydrogenase | 994 | P00356 | 35550 | *Gallus gallus* | 44 | 47 | |
| | Beta enolase | 133 | P07322 | 46839 | *Gallus gallus* | 6 | 17 | |
| | Lactate dehydrogenase A | 100 | P00340 | 36365 | *Gallus gallus* | 2 | 7 | |
| | Creatine kinase | 81 | P00565 | 43301 | *Gallus gallus* | 3 | 12 | |
| | Aldolase | 74 | Q76BC5 | 35908 | *Cephaloscyllium umbratile* | 1 | 3 | R.LAIMENANVLAR.Y |

a

**Table 5. Identification of normalised proteins by 'GeLC-MSMS'.**
For protein identification, 1D gel separations of starting material and beads containing normalised proteins were divided into 22 slices, each of which was digested in-gel with trypsin. Resulting peptide solutions were analysed by LC-ESI-LTQ MSMS and MSMS data were searched against MSDB using MASCOT with the following parameters; taxonomy: Chordata, protease: trypsin, variable modifications: oxidation of methionine, peptide tolerance: 250ppm, MSMS tolerance: 250ppm, peptide charge: 1+, 2+ and 3+, instrument: ESI-TRAP, from which only confident identifications (MOWSE score>45, $p<0.05$) were accepted. Data are presented as follows; a) chicken skeletal muscle soluble proteins, gel slices 1-12, b) chicken skeletal muscle soluble proteins, gel slices 13-22, c) normalised material, gel slices 1-12, d) normalised material, gel slices 13-22. Proteins in red are novel to either sample (before and after normalisation).

| | Chicken skeletal muscle soluble proteins | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gel band # | Protein | MOWSE score | Swissprot accession # | Molecular weight (Da) | First species | # peptides matched | % coverage | Sequence of single peptide match |
| 13 | Glyceraldehyde 3-phosphate dehydrogenase | 266 | P00356 | 35550 | *Gallus gallus* | 4 | 6 | |
| | Creatine kinase | 171 | P00565 | 43301 | *Gallus gallus* | 8 | 21 | |
| | Lactate dehydrogenase A | 127 | P00340 | 36365 | *Gallus gallus* | 6 | 14 | |
| | Beta enolase | 102 | P07322 | 46839 | *Gallus gallus* | 2 | 8 | |
| | Lactate dehydrogenase B | 57 | P42119 | 36358 | *Xenopus laevis* | 1 | 3 | K.KVVDSAYEVIK.L |
| | Phosphoglycerate kinase | 54 | P51903 | 44557 | *Gallus gallus* | 2 | 7 | |
| | Phosphoglucomutase | 49 | Q2UZR2 | 6550 | *Gallus gallus* | 3 | 7 | |
| 14 | Creatine kinase | 247 | P00565 | 43301 | *Gallus gallus* | 10 | 21 | |
| | Glyceraldehyde 3-phosphate dehydrogenase | 213 | P00356 | 35550 | *Gallus gallus* | 8 | 24 | |
| | Beta enolase | 98 | P07322 | 46839 | *Gallus gallus* | 3 | 8 | |
| | Aldolase | 80 | Q76BC5 | 35908 | *Cephaloscyllium umbratile* | 1 | 3 | R.LAIMENANVLAR.Y |
| 15 | Glyceraldehyde 3-phosphate dehydrogenase | 134 | P00356 | 35550 | *Gallus gallus* | 4 | 15 | |
| | Triose phosphate isomerase | 121 | P00940 | 26527 | *Gallus gallus* | 4 | 15 | |
| | Beta enolase | 107 | P07322 | 46839 | *Gallus gallus* | 5 | 10 | |
| | Phosphoglycerate mutase | 95 | Q5ZLN1 | 28749 | *Gallus gallus* | 9 | 35 | |
| | Aldolase | 76 | Q76BC5 | 35908 | *Cephaloscyllium umbratile* | 1 | 3 | R.LAIMENANVLAR.Y |
| | Phosphoglucomutase | 71 | Q2UZR2 | 6550 | *Gallus gallus* | 2 | 4 | |
| | Creatine kinase | 55 | P00565 | 43301 | *Gallus gallus* | 3 | 4 | |
| 16 | Triose phosphate isomerase | 638 | P00940 | 26527 | *Gallus gallus* | 37 | 43 | |
| | Creatine kinase | 185 | P00565 | 43301 | *Gallus gallus* | 4 | 13 | |
| | Titin | 134 | Q9IAR9 | 242090 | Gallus gallus | 5 | 1 | |
| | Glutathione s-transferase | 122 | P20136 | 25488 | *Gallus gallus* | 2 | 10 | |
| | Beta enolase | 89 | P07322 | 46839 | *Gallus gallus* | 2 | 8 | |
| 17 | Adenylate kinase | 203 | P05081 | 21669 | *Gallus gallus* | 14 | 43 | |
| | Beta enolase | 90 | P07322 | 46839 | *Gallus gallus* | 2 | 8 | |
| | Creatine kinase | 87 | P00565 | 43301 | *Gallus gallus* | 4 | 13 | |
| 18 | Phosphatidylethanolamine-binding protein | 70 | P13696 | 20714 | *Bos taurus* | 1 | 7 | K.LYTLVLTDPDAPSR.K |
| | Connectin/titin | 61 | Q9YH41 | 123994 | *Gallus gallus* | 1 | 1 | K.SLVEESQLPEGR.R |
| | Beta enolase | 59 | P07322 | 46839 | *Gallus gallus* | 1 | 2 | R.LITGEQLGEIYR.G |
| | Creatine kinase | 48 | P00565 | 43301 | *Gallus gallus* | 2 | 6 | |
| 19 | Beta enolase | 65 | P07322 | 46839 | *Gallus gallus* | | | |
| 20 | Beta enolase | 62 | P07322 | 46839 | *Gallus gallus* | 1 | 2 | R.LITGEQLGEIYR.G |
| | Superoxide dismutase | 48 | P80566 | 15563 | *Gallus gallus* | 1 | 9 | K.DADRHVGDLGNVTAK.G |
| 21 | Beta enolase | 74 | P07322 | 46839 | *Gallus gallus* | 1 | 2 | R.GNPTVEVDLHTAK.G |
| | Fatty acid binding protein | 79 | Q6DRR5 | 14807 | *Gallus gallus* | 1 | 5 | K.LVDTANFDEYMK.A |
| 22 | Ubiquitin | 96 | P62973 | 17738 | *Gallus gallus* | 4 | 22 | |
| | Aldolase A | 85 | P04075 | 39395 | *Homo sapiens* | 2 | 4 | |
| | Thioredoxin | 75 | P08629 | 11562 | *Gallus gallus* | 1 | 12 | K.SVGNLADFEAELK.A |
| | Creatine kinase | 72 | P00565 | 43301 | *Gallus gallus* | 4 | 13 | |
| | Beta enolase | 62 | P07322 | 46839 | *Gallus gallus* | 1 | 2 | R.GNPTVEVDLHTAK.G |

b

| | Normalised material | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gel band # | Protein | MOWSE score | Swissprot accession # | Molecular weight (Da) | First species | # peptides matched | % coverage | Sequence of single peptide match |
| 1 | Chicken gizzard actin binding protein | 72 | Q90WF0 | 280321 | *Gallus gallus* | 2 | 1 | |
| 2 | Myomesin | 231 | Q90724 | 182054 | *Gallus gallus* | 14 | 8 | |
| | Actin binding protein | 178 | Q90WF0 | 280324 | *Gallus gallus* | 7 | 2 | |
| | Myofibrillar protein | 66 | Q02173 | 309091 | *Gallus gallus* | 1 | 1 | K.SSEISEPVFVEASPGTK.E |
| 3 | Heat shock protein 90 | 143 | Q2TFN9 | 94319 | *Canis familiaris* | 4 | 4 | |
| | Myomesin | 83 | Q90724 | 182054 | *Gallus gallus* | 3 | 2 | |
| | Actin binding protein | 66 | Q90WF0 | 280321 | *Gallus gallus* | 2 | 1 | |
| 4 | Aminopeptidase | 329 | P55786 | 103211 | *Homo sapiens* | 24 | 13 | |
| | Actin binding protein | 62 | Q90WF0 | 280321 | *Gallus gallus* | 6 | 2 | |
| | Heat shock protein 90 | 48 | Q2TFN9 | 94319 | *Canis familiaris* | 2 | 2 | |
| 5 | Molecular chaperone | 482 | P11021 | 72071 | *Homo sapiens* | 21 | 23 | |
| | Immunoglobulin gamma | 93 | S00390 | 53581 | *Gallus gallus* | 2 | 6 | |
| | Heat shock protein 70 | 83 | Q3MHM4 | 72004 | *Bos taurus* | 4 | 8 | |
| | Aminopeptidase | 84 | P55786 | 103211 | *Homo sapiens* | 2 | 2 | |
| | Vinculin | 56 | P12003 | 124560 | *Gallus gallus* | 2 | 1 | |
| 6 | Heat shock protein 70 | 860 | Q75PJ4 | 70827 | *Numidia meleagris* | 61 | 34 | |
| | Immunoglobulin gamma | 102 | S00390 | 53581 | *Gallus gallus* | 3 | 6 | |
| | Albumin | 88 | P19121 | 69918 | *Gallus gallus* | 2 | 4 | |
| 7 | Phosphoglucomutase 1 | 98 | QZUZR2 | 66550 | *Gallus gallus* | 6 | 13 | |
| | Pyruvate kinase | 98 | P00548 | 57978 | *Gallus gallus* | 3 | 4 | |
| | Collapsin response mediator | 98 | Q71SG1 | 62220 | *Gallus gallus* | 10 | 20 | |
| | Disulphide isomerase | 88 | P09102 | 57374 | *Gallus gallus* | 3 | 5 | |
| | Hemopexin | 62 | Q90WR3 | 29366 | *Gallus gallus* | 4 | 3 | |
| | Dystrophin | 51 | S02041 | 422618 | *Gallus gallus* | 1 | 0 | R.SLDLNSIIAEVK.A |
| | Heat shock protein 70 | 49 | O73885 | 70783 | *Gallus gallus* | 3 | 6 | |
| 8 | Glucose 6-phosphate isomerase | 1140 | Q5ZMU3 | 62177 | *Gallus gallus* | 74 | 33 | |
| | Adenysuccinate lyase | 247 | Q5U7AZ | 54606 | *Gallus gallus* | 15 | 18 | |
| | Disulphide isomerase | 246 | P09102 | 57374 | *Gallus gallus* | 10 | 16 | |
| | Phosphoglucose isomerase | 246 | Q8QFU1 | 62057 | *Brachydanio rerio* | 24 | 9 | |
| | Calreticulin | 84 | Q6EE32 | 46851 | *Gallus gallus* | 6 | 23 | |
| 9 | Rab-GDP dissociation inhibitor | 155 | O93382 | 50651 | *Gallus gallus* | 7 | 14 | |
| | Elongation factor 1 | 136 | Q57KM2 | 50153 | *Gallus gallus* | 8 | 8 | |
| | Beta enolase | 118 | P07322 | 46839 | *Gallus gallus* | 6 | 13 | |
| | HSC70 interacting protein | 98 | Q5ZLF0 | 40158 | *Gallus gallus* | 4 | 9 | |
| | Osteoglycin | 66 | Q9W6H0 | 33179 | *Gallus gallus* | 1 | 4 | K.LLLLEELSLAENR.L |
| | Myosin light chain kinase 2 | 61 | Q7LZ16 | 87160 | *Gallus gallus* | 3 | 3 | |
| | Bleomycin hydrolase | 105 | Q6GL32 | 52656 | *Gallus gallus* | 3 | 5 | |
| 10 | Phosphoglycerate kinase | 453 | P51903 | 44557 | *Gallus gallus* | 39 | 23 | |
| | Creatine kinase | 196 | P00565 | 43301 | *Gallus gallus* | 18 | 32 | |
| | Glucose 6-phosphate isomerase | 150 | Q5ZMU3 | 62177 | *Gallus gallus* | 8 | 14 | |
| | Actin A | 112 | P68135 | 40593 | *Oryctolagus cuniculus* | 6 | 21 | |
| | Osteoglycin | 90 | Q9W6H0 | 33179 | *Gallus gallus* | 4 | 13 | |
| | Gelsolin | 82 | O93510 | 85832 | *Gallus gallus* | 2 | 4 | |
| | Beta enolase | 53 | P07322 | 46839 | *Gallus gallus* | 2 | 8 | |
| | Citrate synthase | 46 | P23007 | 48035 | *Gallus gallus* | 1 | 3 | K.GLIYETSVLDPDEGIR.F |
| | Glucose 6-phosphate isomerase | 96 | Q5ZMU3 | 62177 | *Gallus gallus* | 10 | 22 | |
| | Osteoglycin | 91 | Q9W6H0 | 33179 | *Gallus gallus* | 2 | 7 | |
| 11 | Aldolase | 88 | Q76BC5 | 35908 | *Cephaloscyllium umbratile* | 2 | 3 | |
| | Aminopeptidase | 71 | P55786 | 103211 | *Homo sapiens* | 2 | 2 | |
| | Glyceraldehyde 3-phosphate isomerase | 81 | P00356 | 35739 | *Gallus gallus* | 1 | 4 | K.LVSWYDNEFGYSNR.V |
| | Beta enolase | 59 | P07322 | 46839 | *Gallus gallus* | 1 | 2 | R.LITGEQLGEIYR.G |
| 12 | Tropomyosin B | 306 | P19352 | 35746 | *Gallus gallus* | 17 | 36 | |
| | Glyceraldehyde 3-phosphate isomerase | 229 | P00356 | 35739 | *Gallus gallus* | 8 | 21 | |
| | Tropomyosin A | 220 | P04268 | 32692 | *Gallus gallus* | 12 | 27 | |
| | Lactate dehydrogenase B | 125 | P00337 | 36365 | *Gallus gallus* | 3 | 10 | |
| | Beta enolase | 73 | P07322 | 46839 | *Gallus gallus* | 2 | 8 | |
| | Annexin | 72 | Q6B344 | 75179 | *Gallus gallus* | 5 | 8 | |
| | Lactate dehydrogenase | 212 | P00337 | 36365 | *Gallus gallus* | 3 | 10 | |

c

| | Normalised material | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gel band # | Protein | MOWSE score | Swissprot accession # | Molecular weight (Da) | First species | # peptides matched | % coverage | Sequence of single peptide match |
| 13 | Pyruvate dehydrogenase | 112 | Q3TL86 | 38930 | *Mus musculus* | 2 | 7 | |
| | Glycerol 3-phosphate dehydrogenase | 109 | Q7T1E0 | 38169 | *Brachydanio rerio* | 5 | 7 | |
| | Elongation factor 1 | 66 | Q6EE30 | 49811 | *Gallus gallus* | 2 | 3 | |
| 14 | Glycerol 3 phosphate dehydrogenase | 74 | Q6P824 | 38439 | *Brachydanio rerio* | 3 | 7 | |
| | Phosphoglycerate kinase | 62 | P51903 | 44688 | *Gallus gallus* | 4 | 11 | |
| 15 | Tyrosine monooxygenase activator | 205 | Q7ZW20 | 29054 | *Brachydanio rerio* | 9 | 35 | |
| | Protein kinase C inhibitor | 178 | Q5F3W6 | 28213 | *Gallus gallus* | 7 | 27 | |
| | Peroxiredoxin-6 | 75 | Q5ZJF4 | 24977 | *Gallus gallus* | 4 | 18 | |
| | Beta enolase | 50 | P07322 | 46839 | *Gallus gallus* | 1 | 2 | R.LITGEQLGEIYR.G |
| 16 | Connectin/titin | 287 | Q98918 | 464683 | *Gallus gallus* | 20 | 0 | |
| | Apolipoprotein A1 | 130 | P08250 | 30661 | *Gallus gallus* | 7 | 22 | |
| | Ubiquitin hydrolase | 119 | Q9PW67 | 26298 | *Gallus gallus* | 6 | 21 | |
| | Heat shock protein 25 | 80 | Q00649 | 21658 | *Gallus gallus* | 4 | 29 | |
| | Beta actin | 66 | Q6RHR8 | 31321 | *Macaca mulatta* | 3 | 11 | |
| | Alpha actin | 62 | P68135 | 40593 | *Oryctolagus cuniculus* | 2 | 10 | |
| | FKSG30 | 62 | Q9BYX7 | 41989 | *Homo sapiens* | 2 | 9 | |
| | Phosphoglycerate kinase | 51 | P51903 | 44688 | *Gallus gallus* | 2 | 5 | |
| | Peroxiredoxin 6 | 61 | Q5ZJF4 | 24961 | *Gallus gallus* | 3 | 11 | |
| 17 | IgE dependent histamine releasing factor | 120 | | 19518 | *Gallus gallus* | 14 | 35 | |
| | Adenylate kinase | 92 | P05081 | 21669 | *Gallus gallus* | 10 | 36 | |
| | Heat shock protein 25 | 66 | Q00649 | 21658 | *Gallus gallus* | 5 | 29 | |
| | Myosin light chain | 60 | P02604 | 20900 | *Gallus gallus* | 1 | 5 | K.ITLSQVGDIVR.A |
| 18 | Troponin C | 250 | P02588 | 18364 | *Gallus gallus* | 6 | 37 | |
| | Protein kinase C inhibitor | 99 | Q5F3W6 | 28213 | *Gallus gallus* | 3 | 12 | |
| | Heat shock protein 25 | 57 | Q00649 | 21658 | *Gallus gallus* | 1 | 6 | K.YTLPPGVEATAVR.S |
| 19 | Troponin C | 395 | P02588 | 18272 | *Gallus gallus* | 12 | 50 | |
| | Myosin light chain | 136 | P02604 | 16179 | *Gallus gallus* | 8 | 50 | |
| | Calmodulin | 94 | Q6R520 | 16704 | *Oreochromis mossambicus* | 11 | 47 | |
| 20 | Calmodulin | 226 | Q6R520 | 16704 | *Oreochromis mossambicus* | 16 | 59 | |
| | PACSIN 3 | 187 | Q1G1I6 | 50711 | *Gallus gallus* | 4 | 5 | |
| | Protein kinase C inhibitor | 120 | Q9I882 | 13750 | *Gallus gallus* | 5 | 28 | |
| | Nebulin | 106 | Q9DEH4 | 276962 | *Gallus gallus* | 5 | 1 | |
| | Glycine cleavage system protein H | 81 | P11183 | 17999 | *Gallus gallus* | 4 | 14 | |
| | Myosin light chain | 76 | P02604 | 16179 | *Gallus gallus* | 1 | 5 | K.ITLSQVGDIVR.A |
| | Adenysuccinate lyase | 49 | Q5U7AZ | 54606 | *Gallus gallus* | 1 | 4 | |
| 21 | PACSIN 3 | 85 | Q1G1I6 | 50711 | *Gallus gallus* | 1 | 3 | R.ALYDYTGQEADELSFK.A |
| | Apolipoprotein A1 | 77 | p08250 | 30661 | *Gallus gallus* | 1 | 5 | K.LADNLDTLSAAAAK.L |
| | Haemoglobin alpha | 77 | P01994 | 15288 | *Gallus gallus* | 2 | 10 | |
| | Vimentin | 66 | Q6PBS2 | 51538 | *Brachydanio rerio* | 3 | 7 | |
| | Actin A | 56 | P68135 | 40593 | *Oryctolagus cuniculus* | 2 | 10 | |
| 22 | Myosin | 162 | P02604 | 16179 | *Gallus gallus* | 8 | 45 | |
| | Apolipoprotein A1 | 113 | P08250 | 30661 | *Gallus gallus* | 1 | 5 | K.LADNLDTLSAAAAK.L |
| | Actin | 67 | Q8MVP7 | 26950 | *Boltenia villosa* | 2 | 16 | |
| | Connectin | 64 | PN0568 | 148490 | *Gallus gallus* | 3 | 1 | |
| | Calmodulin | 63 | P62144 | 16827 | *Anas platyrhynchos* | 5 | 45 | |
| | Acylphosphatase | 60 | P07032 | 11151 | *Gallus gallus* | 2 | 29 | |
| | Nebulin | 56 | Q9DEH4 | 276962 | *Gallus gallus* | 4 | 1 | |
| | AK | 46 | P07032 | 21669 | *Gallus gallus* | 1 | 1 | K.ATEPVIAFYK.G |

**d**

incubation of 50mg protein with 20mg beads as a new band of the same molecular weight (Figure 76; ~62kDa). This was confirmed by repeated in-gel digestion and LC-MSMS analysis (results not shown).

To quantify the degree of normalisation of chicken skeletal muscle soluble proteins using the QconCAT approach for absolute quantification, QconCAT protein was added to starting material and beads containing normalised protein prior to digestion with trypsin. To ensure complete digestion of proteins bound to Equalizer™ beads, vital when using surrogate peptides for absolute quantification, digested protein was analysed by 1D SDS-PAGE to monitor the complete removal of intact proteins (Figure 79). For absolute quantification, peptides were analysed by LC-ESI-Q-ToF MS using relative signal intensity of analyte ('light') and internal standard ('heavy') peaks. This was expressed as nmol/g tissue before and after normalisation for beads exposed to 50mg and 100mg of protein. This does not provide quantitative, biological data but demonstrates in absolute terms the degree of protein normalisation when both 50mg and 100mg of protein were exposed to the ligand library (Figure 80). The extent to which the most abundant proteins, for example glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and beta enolase (β eno) have been diluted during normalisation is quite striking, being reduced from 1340nmol to 20nmol and 420nmol to 24nmol. A couple of proteins have been enriched; tropomyosin A (TMA) and actin polymerisation inhibitor (API) as these were previously quantified as below the limit of detection (Figure 81). The identity of the peptides used for absolute quantification of these two proteins was confirmed by MSMS (Figure 82) where both 'heavy' and 'light' peptides for API co-eluted at 44.30min and were selected for fragmentation at 44.40 and 44.42 minutes respectively. Both chromatographic peaks were combined to give MSMS spectra for both doubly charged ions and *de novo* sequencing of both confirmed sequence identity, thus normalisation to increase abundance of API in this sample was genuine (Figure 83). The y-ion series for both 'light' and 'heavy' peptide ions also confirms the [$^{13}C_6$]arg/[$^{13}C_6$]lys-labelling with a small peak 6Da heavier clearly observed for y-ions 9-13 for the light isotope, and 6Da lighter for the heavy isotope.

Using the QconCAT protein to quantify the degree of normalisation achieved using Equalizer™ beads has demonstrated, in absolute amounts how abundant proteins that previously dominated analyses by 1D SDS-PAGE and mass spectrometry were reduced with the majority remaining unbound to the beads. This has simultaneously enriched low abundance proteins that were previously not identified by gel electrophoresis, with some proteins having a greater

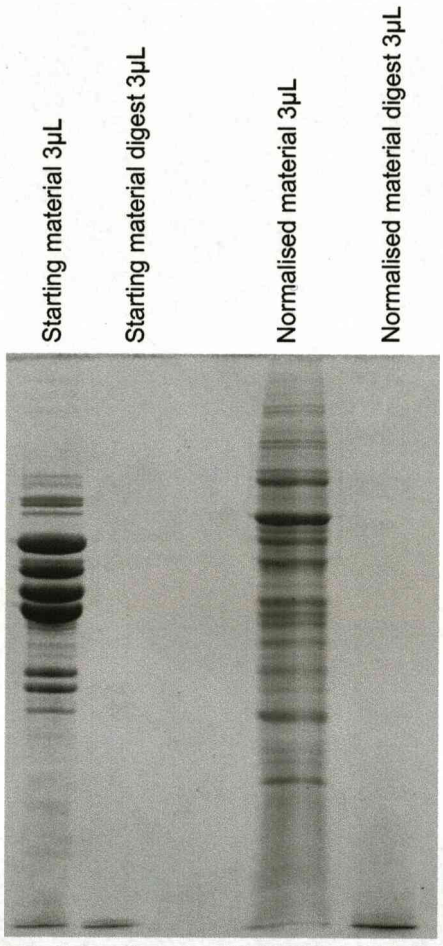



**Figure 79. Proteolysis of chicken skeletal muscle soluble proteins and normalised material.**

100mg soluble protein from chicken skeletal muscle was normalised using Equalizer™ beads. Chicken skeletal muscle soluble proteins and normalised material bound to the ligand library were subjected to proteolysis with trypsin in-solution at an approximate ratio of protein: trypsin of 20:1. Digestion was allowed to proceed for 24h at 37°C after which material before and after normalisation was analysed by 1D SDS-PAGE.

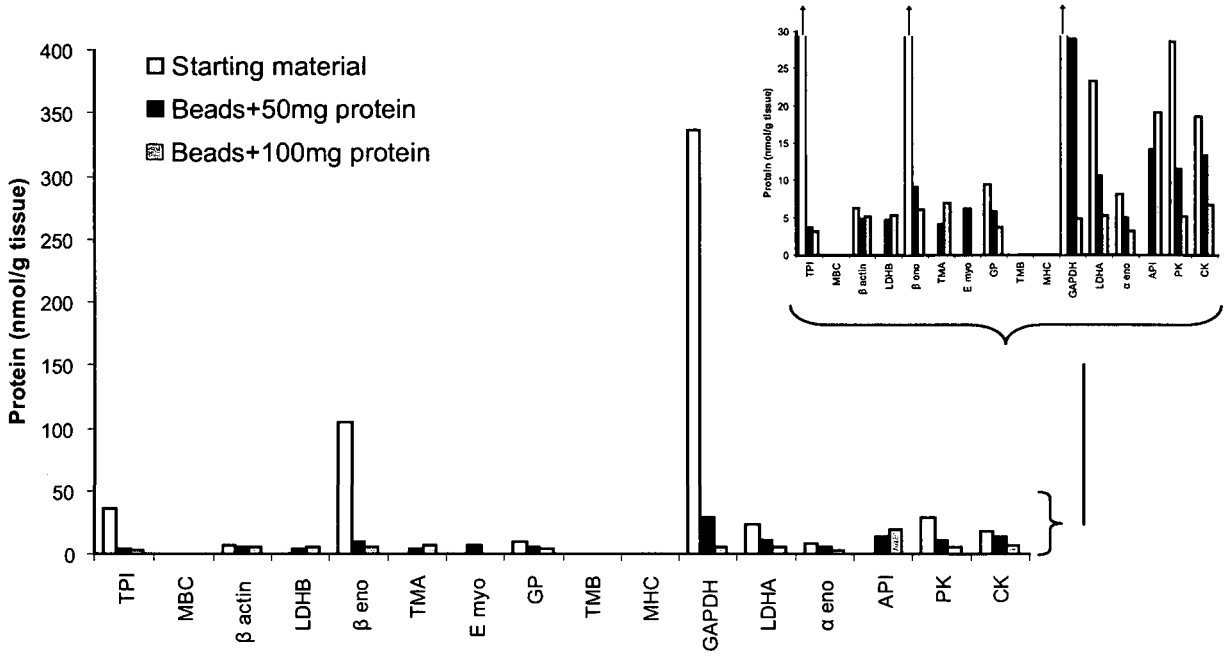



**Figure 80. Quantification of normalisation of chicken skeletal muscle soluble proteins using QconCAT.**

Starting material and beads containing normalised protein (5µL) were diluted 1:10 with 50mM ammonium bicarbonate, to which QconCAT protein was added (starting material+150pmol, beads+36pmol). Protein was digested with trypsin at a ratio of protein: enzyme of 20:1 with incubation at 37°C for 24h after which the digest was incubated with additional trypsin (20:1 substrate:protease) to ensure complete digestion. Peptides were analysed by LC-ESI-Q-ToF MS with the relative intensity of analyte (light; L) and standard (heavy; H) peaks used for absolute quantification of several proteins. Amount of protein present before (white bars) and after normalisation of 50mg protein (black bars) and 100mg protein (grey bars) using 20mg Equalizer™ beads is expressed as nmol/g tissue. The inserted graph to the top right of the main data set are the same data highlighting the y axis from 0-30nmol/g.

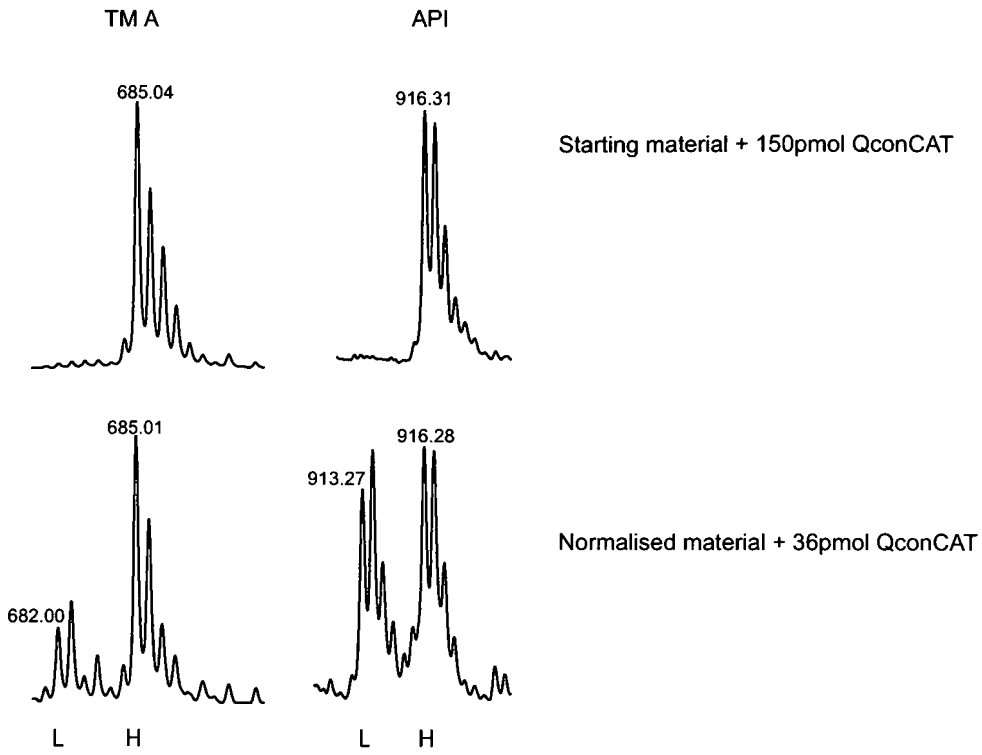



**Figure 81. Analyte and internal standard peptide pairs for two proteins normalised using Equalizer™ beads.**

For quantification of normalisation, starting material and beads containing normalised protein (5µL) were diluted 1:10 with 50mM ammonium bicarbonate, to which QconCAT protein was added (starting material+150pmol, beads+36pmol). Protein was digested with trypsin at a ratio of protein: enzyme of 20:1 with incubation at 37°C for 24h after which the digest was incubated with additional trypsin (20:1 substrate:protease) to ensure complete digestion. Peptides were analysed by LC-ESI-Q-ToF MS with the relative intensity of analyte (light; L) and standard (heavy; H) peaks used for absolute quantification. For two proteins, tropomyosin A (TM A) and actin polymerisation inhibitor (API), mass spectra illustrate the increase in signal intensity of the analyte peptide after normalisation.

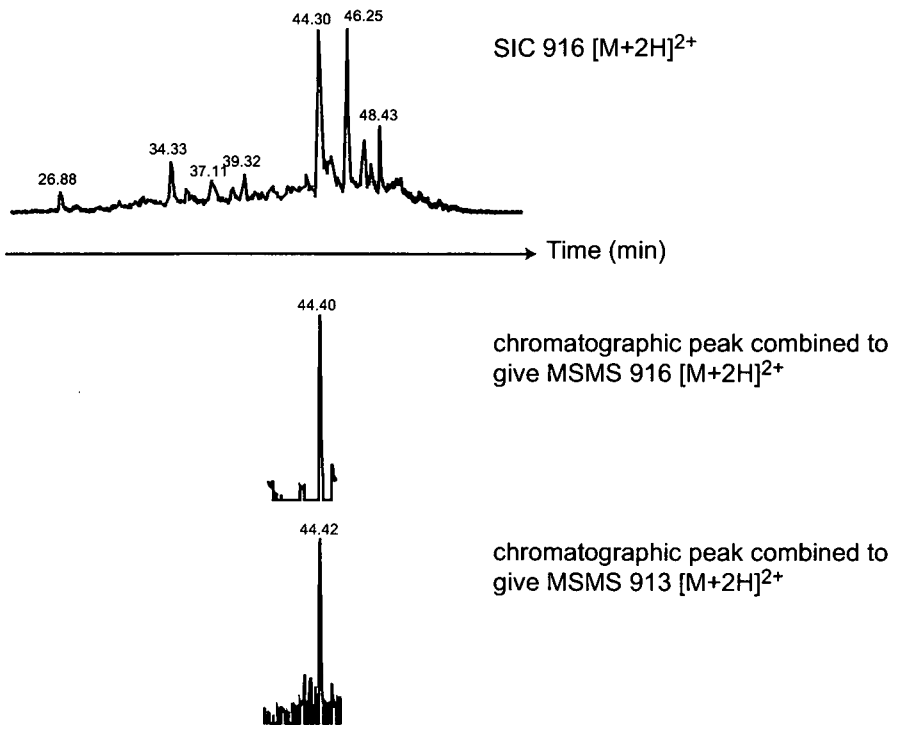



**Figure 82. Chromatographic confirmation of increase in abundance of API with normalisation using Equalizer™ beads.**
Peptides from in-solution digestion with trypsin of normalised chicken skeletal muscle soluble proteins with QconCAT protein were analysed by LC-ESI-Q-ToF MSMS using a collision energy of 30% for fragmentation. A selected ion chromatogram for the internal standard peptide at 916m/z ($[M+2H]^{2+}$) revealed a peak at 44.30min. Chromatograms for MSMS data acquisition also contained a peak at 44.40min and 44.42min which when combined gave rise to mass spectra for MSMS of co-eluting internal standard peptide 916m/z ($[M+2H]^{2+}$) and analyte peptide 913m/z ($[M+2H]^{2+}$).

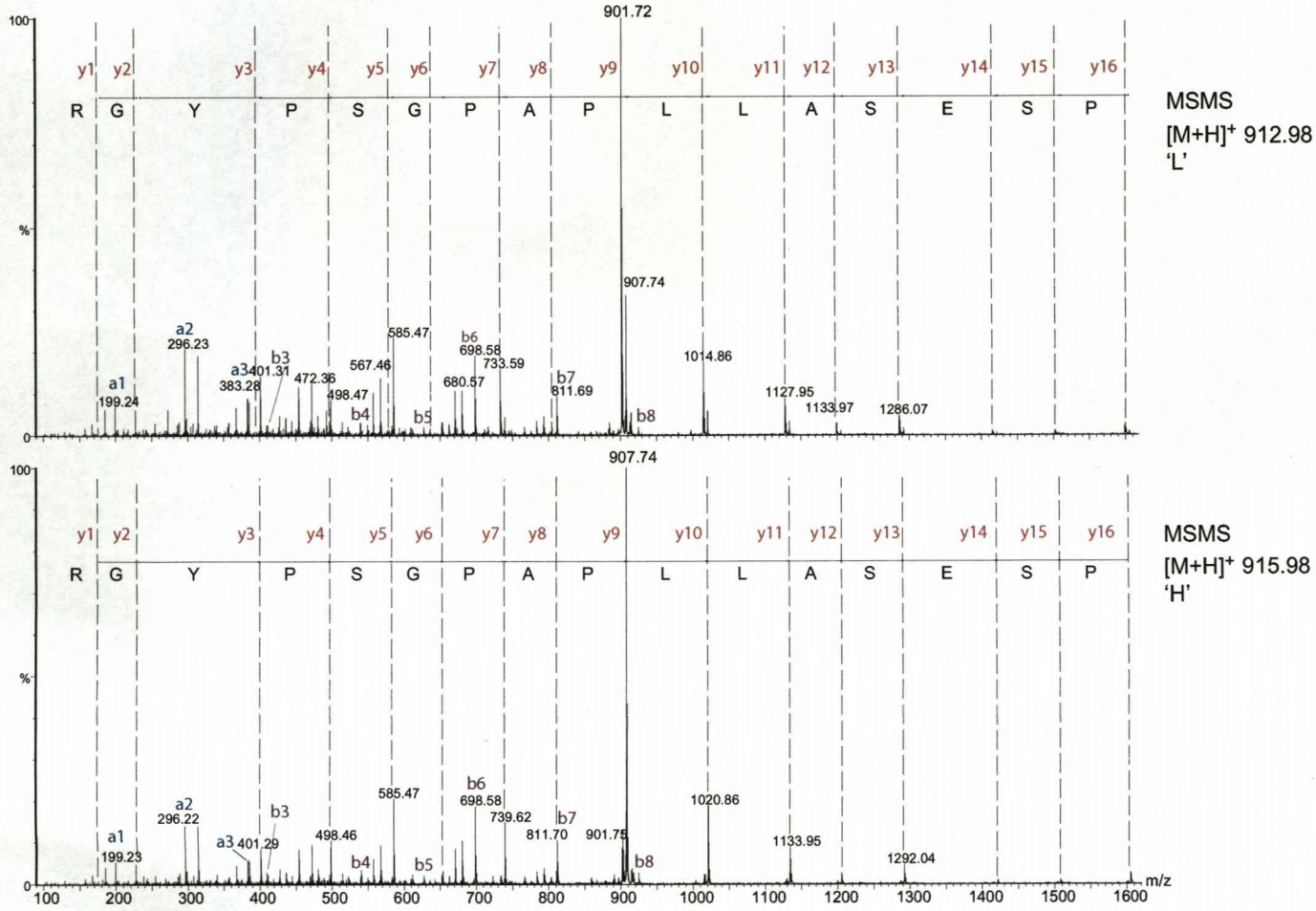



**Figure 83. Confirmation of increase in abundance of API with normalisation using Equalizer™ beads.**
Tandem mass spectra from fragmentation of co-eluting analyte; 913m/z ($[M+2H]^{2+}$) and internal standard peptide; 916m/z ($[M+2H]^{2+}$) from actin polymerisation inhibitor (API). Spectra were interpreted manually to assign the amino acid sequence for each peptide.

affinity for the ligand library than others, resulting in greater enrichment, or less reduction in abundance and as such, all proteins are not normalised to the same level. This technology is not applicable to quantitative analysis of proteomes; however, it will provide a very useful tool for protein identification of complex mixtures exhibiting dramatic dynamic range between high and low abundance proteins.

### 5.6.3 Absolute quantification of the post-translational modification, deamidation

From analysis of chicken skeletal muscle soluble proteins, one of the most abundant is glyceraldehyde 3-phosphate dehydrogenase (GAPDH), amounting to 11 ± 1% (mean ± SEM, n=3) of soluble protein when resolved by 1D SDS-PAGE (Figure 31) and analysed by densitometry, and up to 500 ± 50 nmol/g (mean ± SEM, n=4) tissue when analysed using the QconCAT method for absolute quantification. MALDI-ToF mass spectra for this protein, digested in-gel with trypsin prior to MS analysis were of high quality and gave very high probability identification of this protein (Table 4, supplementary figure 18). Close inspection of each peptide in the mass spectrum indicated that for most, the observed mass isotopomer distribution was as expected, and was in close agreement to the distribution predicted by the MsIsotope program (http://prospector.ucsf.edu/). One peptide in particular (VKVGVNGFGR, $[M+H]^+$ 1032.58m/z) was notably different from the others; the isotope distribution profile was not as predicted (Figure 84). In particular, the relative intensity of the monoisotopic ion (1032.59m/z) was diminished, and of lower intensity than the first [$^{13}C$] isotopomer (1033.59m/z); a relative intensity pattern that is unexpected for a peptide of mass 1031.58Da, given an empirical formula of $C_{46}H_{78}N_{15}O_{12}$.

The mass isotopomer envelope is consistent with the analyte being a mixture of two peptides, one of monoisotopic 1032.58m/z and a second at a monoisotopic of 1033.58m/z. The higher m/z peptide could have been a contaminant or it could have been generated from the peptide at 1032.58m/z. If so, the most probable explanation for the mass increase was deamidation of the asparagine residue, which, by conversion to an aspartate residue would increase the mass by 0.985Da (-$NH_2$ to -OH). To prove that the atypical profile was a consequence of deamidation, the peptide mixture was esterified by reacting with acetyl chloride and methanol to convert carboxyl groups to their methyl esters, resulting in a mass shift of 14.03Da. As the peptide **$V_2$KVGVNGFGR$_{10}$** would possess a single carboxyl group in the amide form (the alpha carboxyl group), and two in the acid form, deconvolution of the atypical peptide into two products, one esterified at a single position (+14.03Da), and a second modified in two positions

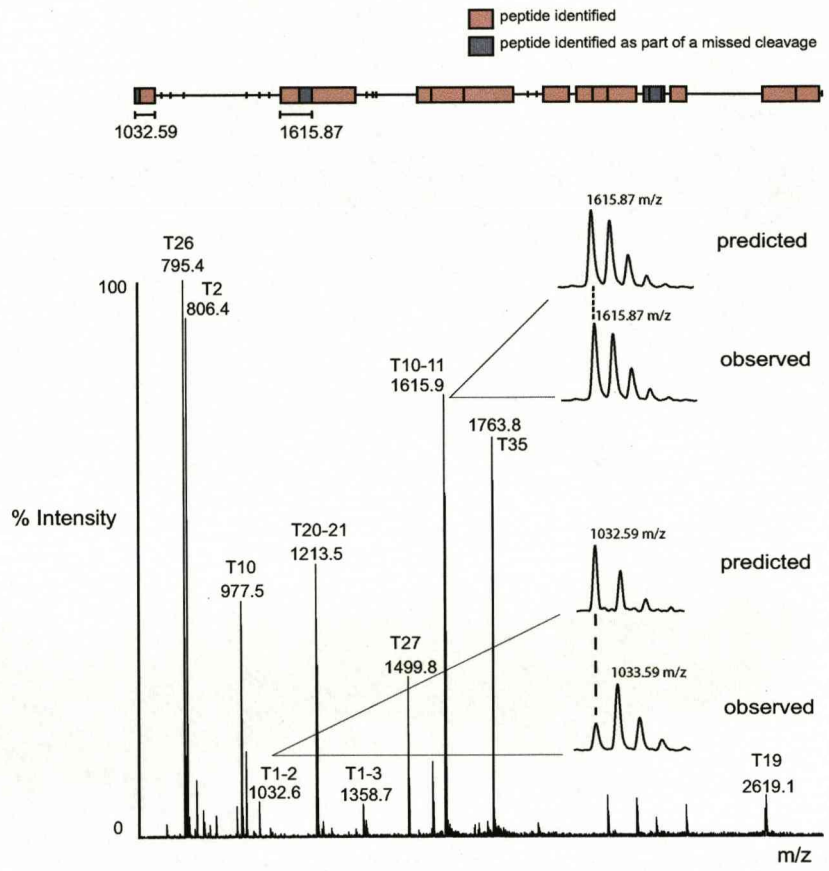



**Figure 84. Atypical peptide mass spectrum consistent with deamidation.**
Glyceraldehyde 3-phosphate dehydrogenase (GAPDH; 1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) purified from rabbit skeletal muscle (Sigma, Dorset, UK) was digested in-solution with trypsin at a substrate:protease ratio of 50:1 by weight, and the masses of the resultant tryptic peptides were assessed by MALDI-ToF MS; a coverage map is included at the top of the figure with peptides identified represented in pink and those identified as part of a missed cleavage in blue. The spectrum of a typical partial cleavage tryptic peptide (T10-11, m/z 1615.9; indicated on the peptide map) was compared with the mass spectrum predicted by the MS-Isotope tool (http://prospector.ucsf.edu/). This behaviour, common to almost all other peptides, emphasised the atypical profile observed for the N-terminal partial cleavage peptide (T1-2, m/z 1032.6).

(+28.06Da) upon esterification would confirm deamidation. When the peptide mixture was analysed after esterification, the MALDI-ToF MS ions 1032/1033m/z disappeared, and two new ions appeared, one representing the single modified amide (m/z 1032.58+14.03 = 1046.61) and the second reflecting the double modified acid (m/z 1033.58+28.06 = 1061.64; Figure 85, this analysis was conducted by L. McDonald). This confirmed deamidation of the asparagine residue, leading to further investigation, particularly for the impact this might have on protein identification, and quantification.

To assess the extent of deamidation in the native protein, and ascertain whether the residue had deamidated *in vivo*, or was an artifact of sample preparation and processing, purified rabbit GAPDH (Sigma, Dorset, UK) was digested with trypsin at 37°C over 24h. Proteolysis was stopped at selected time points during the digestion reaction by mixing with 10% (v/v) formic acid, and the resulting peptides were analysed by MALDI-ToF MS. Deamidation was monitored as the partition between acid and amide variants of the peptide in MALDI-ToF mass spectra using peak height data and was plotted as a function of time (Figure 86). The N-terminal peptide of GAPDH (**VKVGVNGFGR**) was released within a few minutes and was readily detected as the first analyte ion to appear in the MALDI-ToF mass spectrum. In the early stages of digestion, the mass spectrum of this peptide was entirely consistent with it being exclusively in the amide form (peptide isotopomer was as predicted). However, as time progressed during proteolysis, the mass spectrum of the peptide showed that the peptide was converted to a mixture of the amide and acid variants, and after 10h of digestion, the peptide was over 80% in the acid form. Proteolysis with the protease Asp-N which cleaves N-terminal to an aspartic acid residue, was also conducted in the same way with the reaction stopped at selected time points during digestion by addition to formic acid. Peptides were analysed by MALDI-ToF MS (AXIMA ToF$^2$, Shimadzu, Manchester, UK) with pulsed extraction immediately proceeding ionisation set to achieve optimal resolution at 2600m/z (Figure 87). After 30min digestion the peptide ion containing the asparagine residue is clear (2602.6m/z; although this is overlapping with other peptide species at higher m/z) but by 24h digestion, this peak has disappeared, reflecting the cleavage of this peptide N-terminal to the aspartic acid residue as this is converted from asparagine by deamidation. Although this experiment does not contribute additional information as to the extent of deamidation, it confirms the process continues after initial proteolytic attack, and is not exclusive to the tryptic release of the smaller N-terminal peptide.

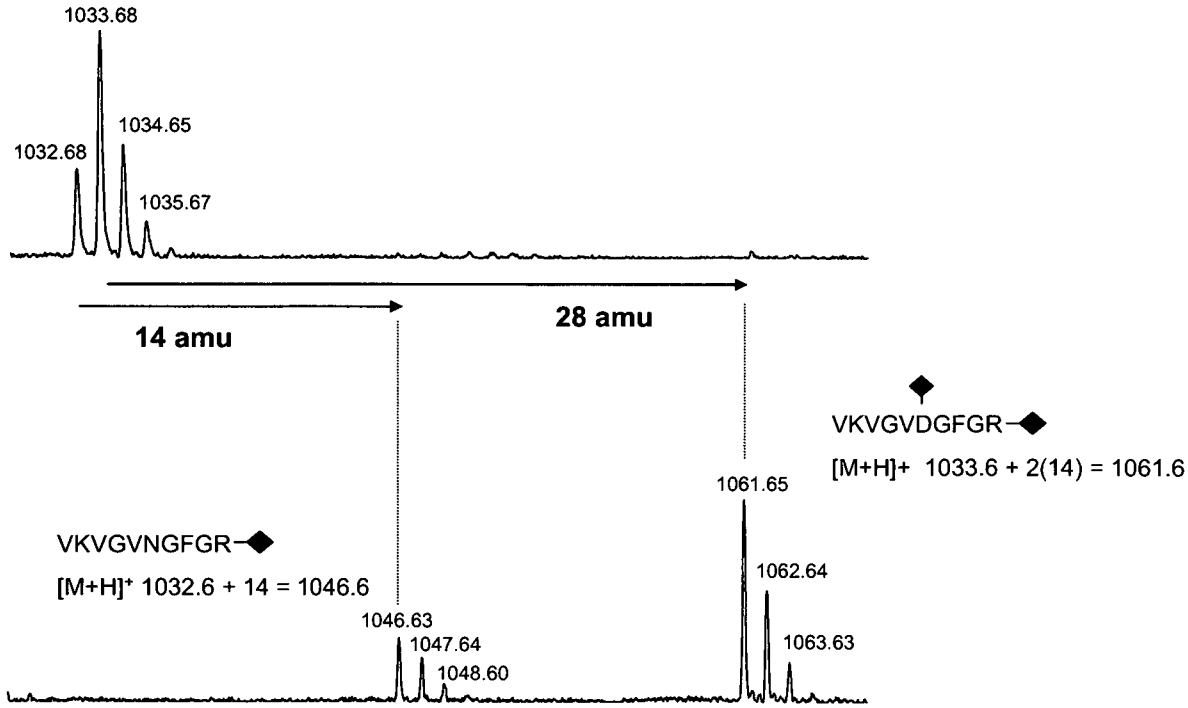



**Figure 85. Esterification of acidic residues in the N-terminal peptide of GAPDH.**
Tryptic peptides recovered from an in-gel tryptic digest of GAPDH (purified from rabbit skeletal muscle, Sigma, Dorset, UK) were reacted with acetyl chloride and methanol to convert acidic residues to their corresponding methyl esters. The upper mass spectrum is the peptide resulting from partial deamidation of Asn6, thus is a mixture of two forms (asparagine containing and aspartic acid containing). The lower spectrum, obtained after esterification has resolved the peptide into two distinct reaction products at 1046.63 m/z and 1061.65 m/z, consistent with the addition of one and two methyl groups (+14.03Da), respectively.

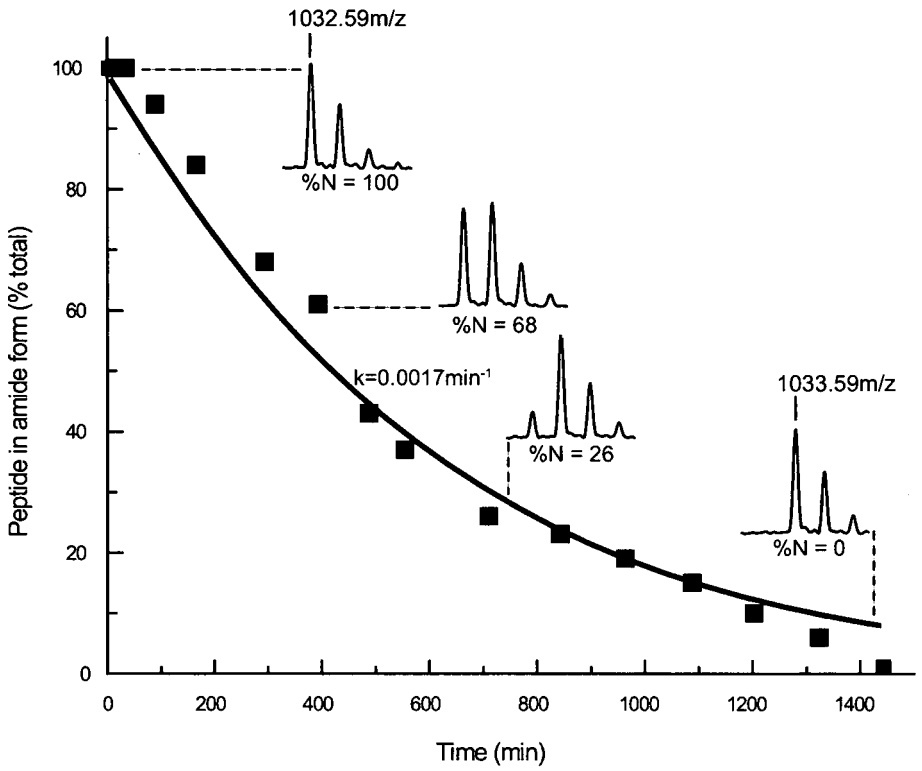



**Figure 86. Time course of deamidation of the N-terminal peptide of GAPDH.**
Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, UK; 1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) was digested with trypsin (trypsin:protein 1:100) over 24h at 37°C. Proteolysis was stopped at 0, 2, 5, 10, 30, 60, 120, 240, 480 and 1440 min by mixing 10µL from the digestion mixture with 10µL 10% (v/v) formic acid. The resulting peptides were analysed by MALDI-ToF MS and deamidation was monitored during proteolysis for the N-terminal peptide of sequence VKVGVNGFGR at 1032.59 m/z. The proportion of acid and amide variants was assessed from peak height data, and plotted as a function of time. Peptide envelopes illustrating the conversion of acid to amide form in MALDI-ToF mass spectra corresponding to time points over 24h are inserted above the data. The solid line is the trajectory taken by first order decay ($y=100e^{-kt}$) for the proteolysed GAPDH.

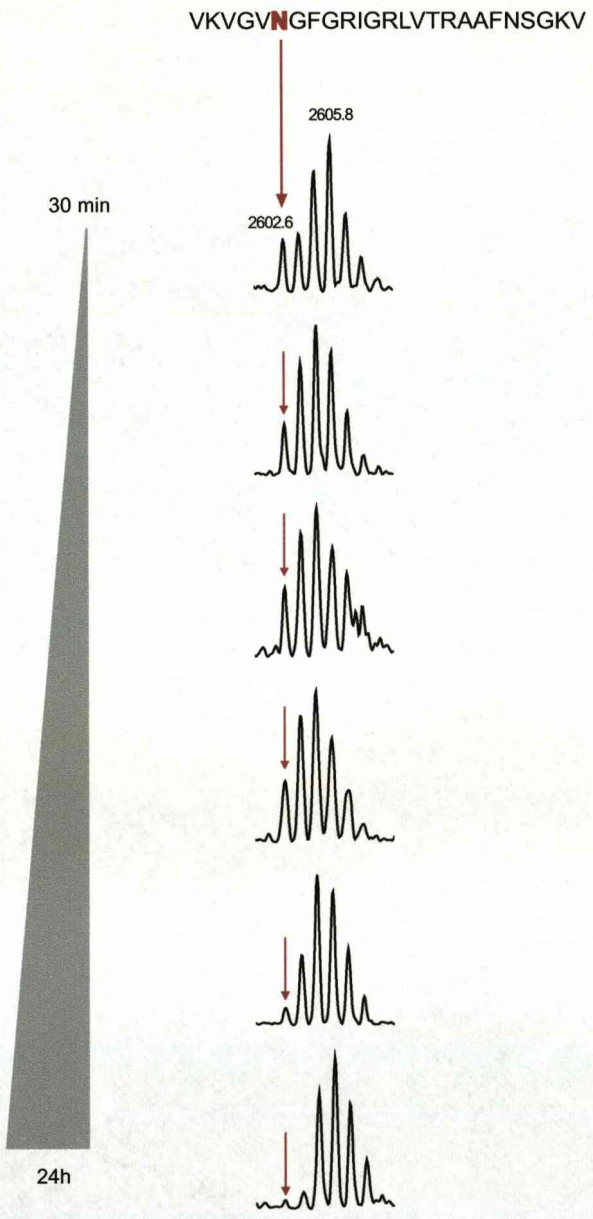



**Figure 87. Time course of deamidation of the N-terminal peptide of GAPDH during proteolysis with Asp-N.**

Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, UK; 1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) was digested with Asp-N (protease:protein 1:100) over 24h at 37°C. Proteolysis was stopped at 0, 2, 5, 10, 30, 60, 120, 240, 480 and 1440 min by mixing 10µL from the digestion mixture with 10µL 10% (v/v) formic acid. The resulting peptides were analysed by MALDI-ToF MS using a laser energy of 85% with pulsed extraction set to achieve optimal resolution at 2600m/z (AXIMA-ToF2, Shimadzu, Manchester, UK). Deamidation was monitored as the disappearance of the monoisotopic peak (VKVGVNGFGRIGRLVTRAAFNSGKV: 2602.6 m/z) as asparagine was converted to aspartic acid causing subsequent cleavage N-terminal to the aspartic acid residue by AspN during proteolysis.

In order to draw conclusions relating to the extent of deamidation, by investigations of the rate of this reaction in the native protein, it was essential to confirm that conversion of asparagine to aspartic acid had no effect on the ionisation of this peptide when analysed by MALDI-ToF MS. For this, the N-terminal peptide of GAPDH (of sequence VKVGVNGFGR and mass 1041.59Da) was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide. A proportion of synthetic peptide was reconstituted in 50mM ammonium bicarbonate prior to incubation at 37°C for 24h to allow complete deamidation of asparagine. A further proportion of synthetic peptide was reconstituted in 10% (v/v) formic acid to prevent deamidation of asparagine. Fully deamidated (100% aspartic acid containing), and non-deamidated (100% asparagine) peptide were mixed in known ratios and analysed by MALDI-ToF MS using peak height data to calculate the proportion of asparagine and aspartic acid (Figure 88). This analysis confirmed that there was no difference in ionisation of asparagine, or aspartic acid containing peptide, with a slope of 0.99 and $R^2$ of 1.0. As such, data as to the relative proportion of acid and amide variants during deamidation could be used to assess the rate of deamidation.

The first order rate constant for deamidation during proteolysis of purified rabbit GAPDH with trypsin (Figure 86) was approximately 0.0017min$^{-1}$, which was higher than the value derived from model peptides – for the sequence $NH_2$**GVNGG**OH the first order rate constant was previously measured at 0.0004min$^{-1}$ (Robinson *et al.,* 2004). However, the buffer conditions for the two experiments are not identical; temperature and pH have a large effect on deamidation rate. To investigate the effect of temperature on deamidation, the chemically synthesised peptide, designed as a stable isotope labelled internal standard for the GAPDH N-terminal peptide, was reconstituted in 50mM ammonium bicarbonate, pH 8.8 and incubated for 24h at 23°C, 37°C and 60°C. During this time, a proportion of each was added to 10% (v/v) formic acid to prevent further deamidation and immediately spotted onto a MALDI target with acidic matrix (α-cyano hydroxycinnamic acid) and air dried. The proportion of acid and amide variant was assessed at each time point from peak height data in MALDI-ToF mass spectra and the proportion amide was plotted against time for incubation at each temperature (Figure 89a). For deamidation at each incubation temperature, first order decay ($y=100e^{-kt}$) was applied to calculate the rate constant. The effect of temperature on deamidation of this model peptide is clear, and to confirm temperature dependence, the data were fitted to the Arrhenius equation whereby a plot of ln(k) against 1/T (temperature, Kelvin) gives a straight line (Figure 89b). To compare the rate of deamidation in the native protein, with that achieved at the same

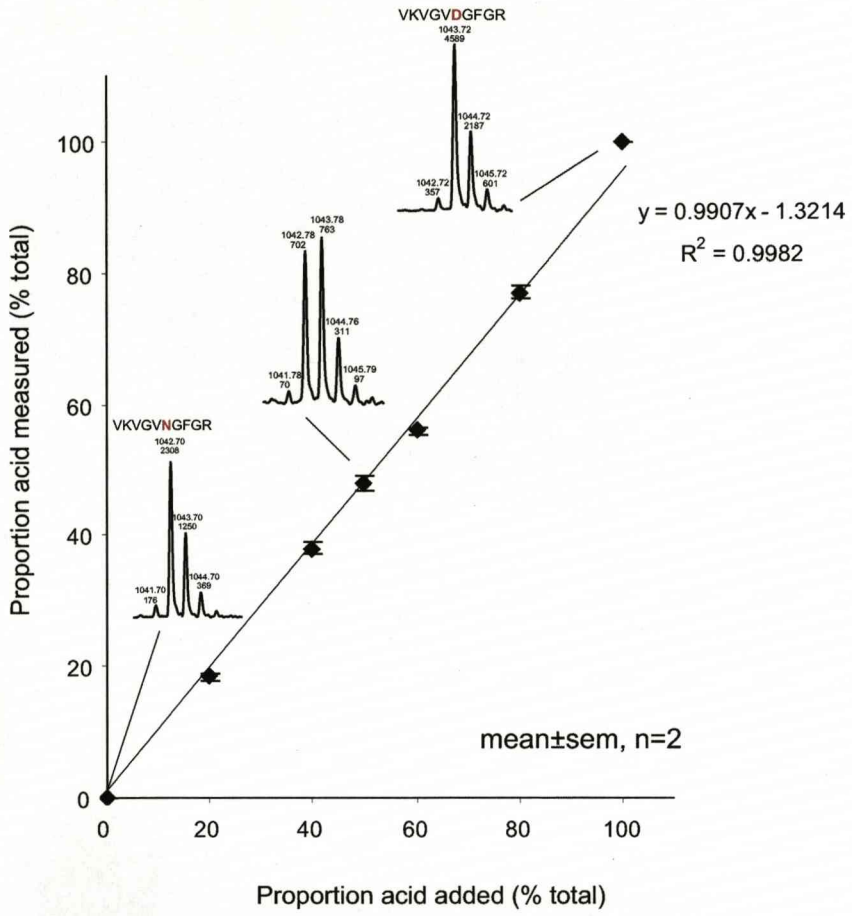
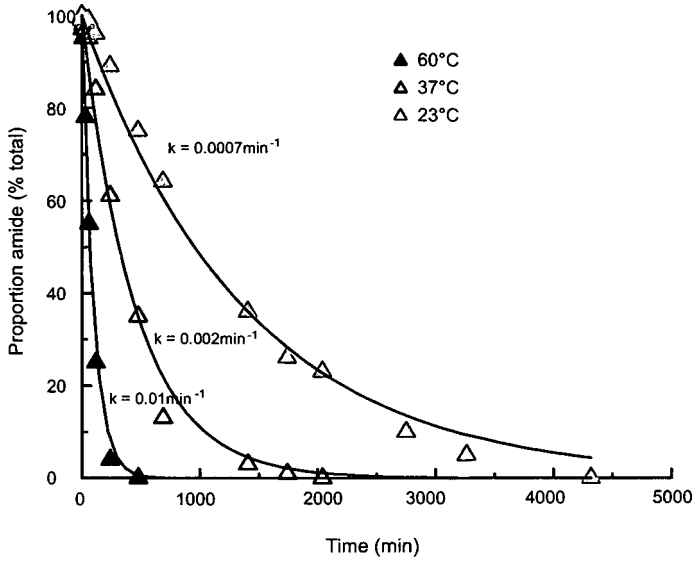



**Figure 88. Ion signal response in MALDI-ToF MS from asparagine and aspartic acid containing peptide.**

The N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and mass 1041.59Da, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide. 1nmol synthetic peptide was reconstituted in 1mL 50mM ammonium bicarbonate, of which a proportion was incubated at 37°C for 24h permitting complete deamidation of asparagine to aspartic acid. A further 1nmol synthetic peptide was reconstituted in 10% (v/v) formic acid and peptide containing asparagine and that containing aspartic acid were mixed in known ratios (0-100%). The proportion of acid and amide variant was assessed from peak height data and the percentage acid variant added was correlated with the percentage of acid measured when the two were mixed and analysed by MALDI-ToF MS. Data are presented as mean±sem, n=2 with spectra for 100% amide, 50% amide/50% acid and 100% acid presented as inserts.

a




b

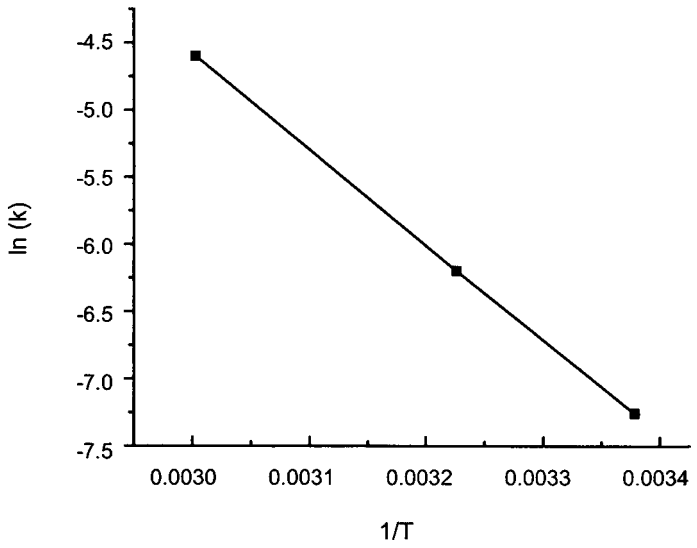



**Figure 89. The effect of temperature on peptide deamidation.**
The N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and mass 1041.59Da, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide. 1nmol synthetic peptide was reconstituted in 1mL 50mM ammonium bicarbonate, of which 3x15µL were diluted 1:10 with 50mM ammonium bicarbonate and incubated at 23°C, 37°C and 60°C for 24h during which time 10µL peptide was added to an equal volume of 10% (v/v) formic acid at selected time points to prevent further deamidation. The proportion of acid and amide variant was assessed at each time point from peak height data upon analysis by MALDI-ToF MS and the percentage amide variant measured was plotted against time for the incubation at each temperature (a). The solid line is the trajectory taken by first order decay ($y=100e^{-kt}$) for the deamidated peptide. To confirm the temperature dependence on the rate of deamidation, the data were fitted to the Arrhenius equation whereby a plot of ln(k) for deamidation at each temperature against 1/T gives a straight line (b).

incubation temperature for the synthetic peptide, proportion of amide was plotted against time for both (Figure 90) with the rate of deamidation for the native protein, 0.0017min$^{-1}$ and that using a synthetic peptide of the same sequence, was 0.0023min$^{-1}$. The higher rate of deamidation of the synthetic peptide is likely to reflect an association between the partially digested protein and the N-terminal peptide, restricting the adoption of a conformation favourable for deamidation, thus diminishing the deamidation rate. However it is clear that the difference between deamidation in a model peptide and that for the native protein, is that the peptide must first be released by proteolysis, and this warranted further investigation into the relationship between these two simultaneous processes.

As the N-terminal peptide itself contains an internal tryptic cleavage site (**VK – VGVDGFGR**), the peptide **VKVGVDGFGR** (summed across acid or amide forms) decreased slowly during proteolysis of the native protein. A model of the sequential first order processes of proteolysis ($k_1$) of the native protein ($N_{native}$) to release the amide form of the peptide (**VKVGVNGFGR**) followed by deamidation ($k_2$) to generate the acid form (**VKVGVDGFGR**) was created (Figure 91). This also incorporates the secondary process of proteolysis ($k_3$) of the N-terminal peptide containing either acid or amide, to release the ValLys dipeptide. Assuming that the rate of deamidation ($k_2$) was independent of the N-terminal ValLys dipeptide and that the rate of tryptic removal of the N-terminal dipeptide ($k_1$) was the same, irrespective of whether the peptide was in acid or amide form, the change in amount (relative to the initial amount of protein,) of the larger peptides (**VKVGVNGFGR** + **VKVGVDGFGR**, *N+D*) as a function of time, is given by:

$$N + D = N_{native(t=0)} \cdot \left( \frac{k_1}{k_3 - k_1} \cdot (e^{-k_1 \cdot t} + e^{-k_3 \cdot t}) \right) \quad \text{.....(1)}$$

As part of the same process, the shortened peptide (**VGVNGFGR** + **VGVDGFGR**, *N'+D'*) appears according to:

$$N' + D' = N_{native(t=0)} \cdot \left( 1 - \frac{k_3}{k_3 - k_1} \cdot e^{-k_1 \cdot t} + \frac{k_3}{k_3 - k_1} \cdot e^{-k_3 \cdot t} \right) \quad \text{.....(2)}$$

To use this model to investigate the kinetics of both deamidation and proteolysis, the stable isotope labelled synthetic peptide, identical to the N-terminal peptide of GAPDH, with a 10Da mass offset relative to the natural peptide due to labelling of arginine with both [$^{13}C_6$] and [$^{15}N_4$],

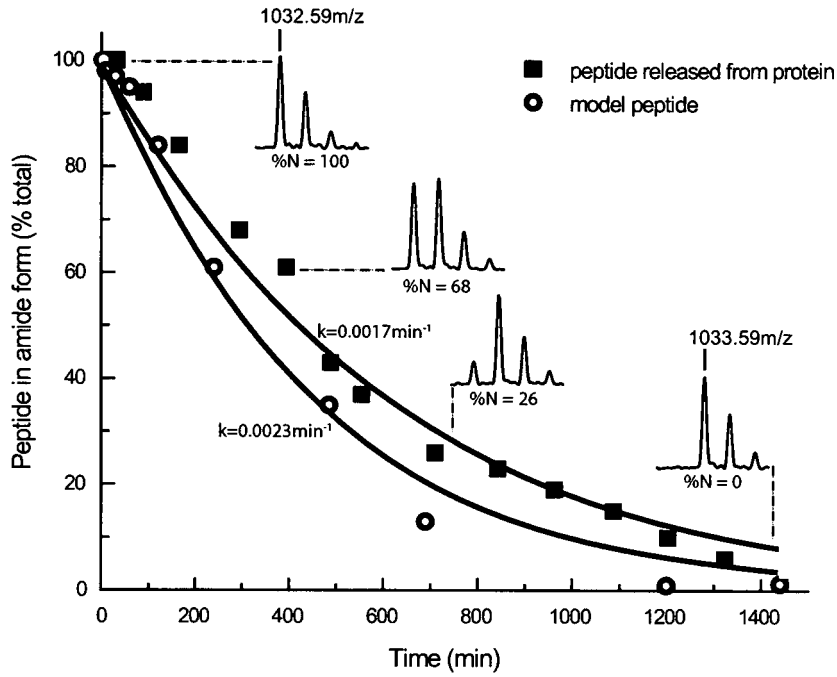



**Figure 90. Time course of deamidation of the N-terminal peptide of GAPDH.**
Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, UK; 1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) was digested with trypsin (trypsin:protein 1:100) over 24h at 37°C. Proteolysis was stopped at 0, 2, 5, 10, 30, 60, 120, 240, 480 and 1440 min by mixing 10µL from the digestion mixture with 10µL 10% (v/v) formic acid. The resulting peptides were analysed by MALDI-ToF MS and deamidation was monitored during proteolysis for the N-terminal peptide of sequence VKVGVNGFGR at 1032.59 m/z. The proportion of acid and amide variants was assessed from peak height data, and plotted as a function of time (closed squares). Peptide envelopes illustrating the conversion of acid to amide form in MALDI-ToF mass spectra corresponding to time points over 24h are inserted above the data. To compare this with model peptide studies, the N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and mass 1041.59Da, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide. This peptide was incubated in 50mM ammonium bicarbonate at 37°C and a sample of the peptide was added to an equal volume of 10% (v/v) formic acid at selected time points. The relative amounts of acid and amide variants of the peptide were measured using MALDI-ToF MS and this was used to calculate the rate of deamidation; these data are presented as open circles. The solid lines are the trajectories taken by first order decay ($y=100e^{-kt}$) for the synthetic peptide and the proteolysed GAPDH.

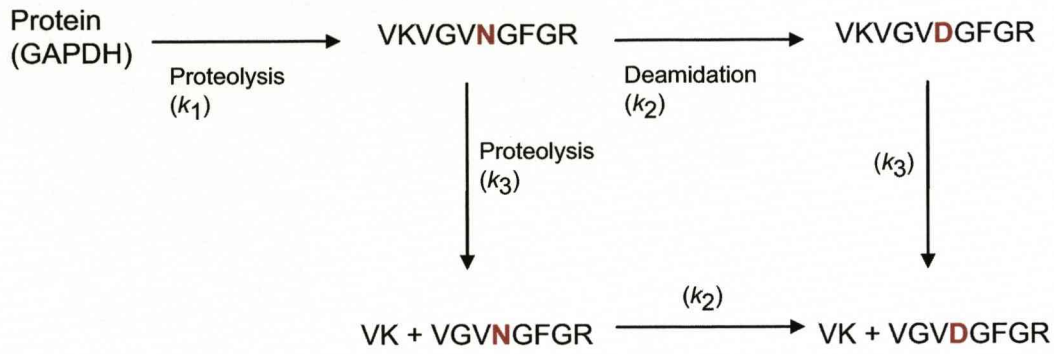



**Figure 91. Model of proteolysis and deamidation of the N-terminal peptide of GAPDH.**
The simultaneous processes of proteolysis and release of the N-terminal peptide of GAPDH followed by deamidation of the asparagine residue to aspartic acid were modelled according to this scheme. The model also included the subsequent proteolysis of the N-terminal peptide (VKVGVNGFGR or VKVGVDGFGR) at the internal lysine residue to generate a dipeptide and a truncated peptide (VK+VGVNGFGR or VK+VGVDGFGR). In this scheme, the rate of deamidation was assumed to be the same, whether in the full length or truncated N-terminal peptide, and that the rate of removal of the N-terminal dipeptide was independent of the amide/acid variants.

was used to monitor the behaviour of the peptide during proteolysis. For quantification (Kirkpatrick *et al.*, 2005), purified rabbit GAPDH was digested with trypsin and digested material was added to a known amount of synthetic peptide in 10% (v/v) formic acid to stop digestion and prevent further deamidation at selected time points. Peptides were analysed by MALDI-ToF MS and the relative intensities of analyte peptide and internal standard were used to quantify the amount of peptide released from the protein, monitoring both the N-terminal peptide and the shorter peptide produced from further proteolysis (Figure 92). As conversion of asparagine to aspartic acid alters the isotope envelope of the analyte peptide, the composite abundance of the entire isotopic envelope for both analyte and internal standard peptide was summed in each case. Assuming that the rate of tryptic cleavage is consistent for both acid and amide variants, from these equations, the second order rate constants were calculated (first order rate constant divided by protease concentration) for initial release of the large peptide ($k_1$) and the rate of proteolysis of this large peptide ($k_3$). The value of $k_1$ was estimated to be $1.22\pm0.025 min^{-1}.\mu M$ and for $k_3$, $0.50\pm0.008 min^{-1}.\mu M$ (trypsin=0.2μM). As expected, the endoproteolytic release of the longer peptide is faster than the release of the N-terminal dipeptide, as trypsin is known to act poorly as a dipeptidyl peptidase. However, the release of the longer peptide is likely to be suppressed by the three dimensional structure of the protein.

To investigate the effects of the higher order structure of GAPDH on proteolysis and subsequent deamidation, we analysed the X-ray crystal structure of rabbit GAPDH (PDB code 1J0X.PDB). First, we used the tool NickPred (Hubbard, 1998) which although designed to predict sites of proteolytic attack, can generate a comprehensive analysis of the environment of every residue in a protein sequence. The N-terminal region of GAPDH is rather constrained, exhibiting low temperature factors (B-values) and low protrusion and accessibility (results not shown). Close inspection of the structure in the vicinity of $Asn_6$ revealed an extensive hydrogen bonded network that might be expected to constrain main chain flexibility and to reduce the propensity for asparagine deamidation (Figure 93). However, once the peptide was released by proteolysis (within the first two minutes of digestion with a low concentration of trypsin, Figure 94), deamidation proceeded at a higher rate than that predicted from model studies. These experiments are consistent with the following propositions; that the residue in the intact protein is exclusively in the amide form, secondly that the tryptic fragment containing the amide residue can undergo deamidation and thirdly, that deamidation is not an artefact of the mass spectrometric analysis. Excision of the peptide from the GAPDH structure relieves the constraint in the peptide backbone trajectory, permitting the deamidation reaction to take place.

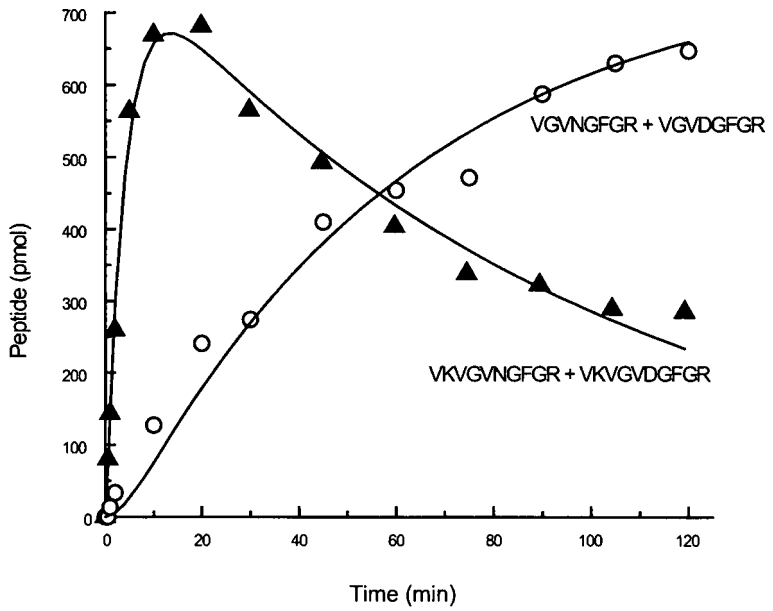



**Figure 92. Absolute quantification of proteolysis of the GAPDH N-terminus.**
Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, UK; 1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) was digested with trypsin (trypsin:protein 1:10) over 24h at 37°C. The N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and mass 1041.59Da, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide. For quantification of proteolysis, the synthetic peptide was added to digested material in 10% (v/v) formic acid to stop digestion at selected time points. Peptides were analysed by MALDI-ToF MS and the relative intensities of analyte peptide and internal standard were used to quantify the amount of peptide released from the protein during incubation with trypsin at 37°C. Both the N-terminal peptide (VKVGVNGFGR/VKVGVDGFGR; 1032.59 $[M+H]^+$; closed triangles) and the shorter peptide produced by further proteolysis (VGVNGFGR/VGVDGFGR; 805.59 $[M+H]^+$; open circles) were monitored. As conversion of asparagine to aspartic acid alters the isotope envelope of the analyte peptide, the composite abundance of the entire isotopic envelope for both analyte and internal standard peptide was summed in each case. The solid lines reflect the fitted curves for the transient appearance of the N-terminal peptide (VKVGVNGFGR/VKVGVDGFGR) and the truncated product (VGVNGFGR/VGVDGFGR), modelled and fitted as sequential first order reactions (see text).

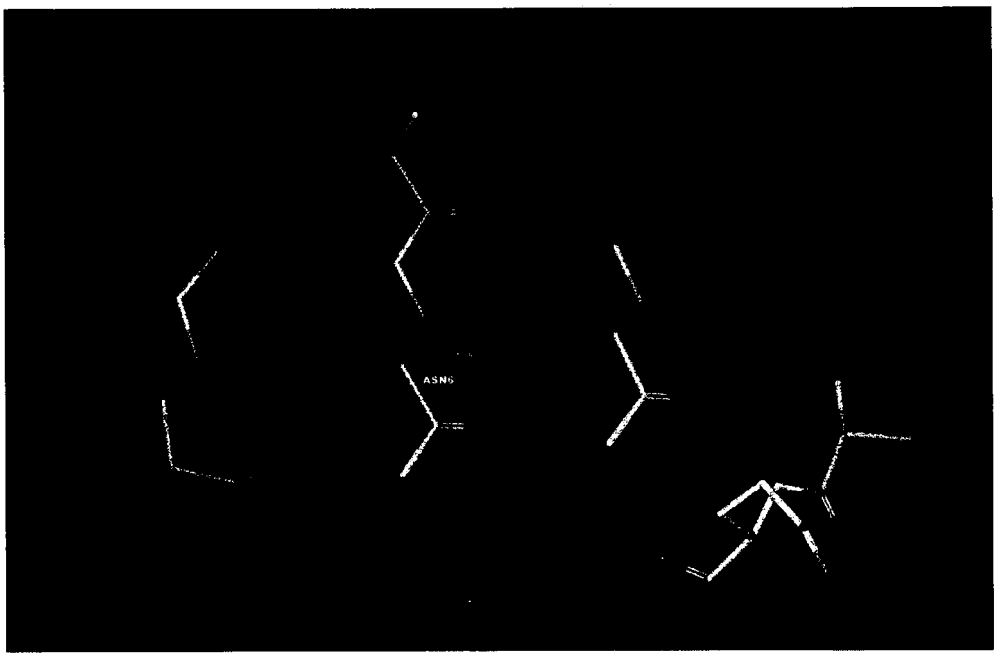



**Figure 93. 3D structure of rabbit skeletal muscle GAPDH .**
X-ray crystal structure of the N-terminal region of rabbit skeletal muscle GAPDH (PDB code 1J0X) highlighting the Asn6Gly7 deamidation site and the local hydrogen bonded environment. The green dashed lines denote hydrogen bonds.

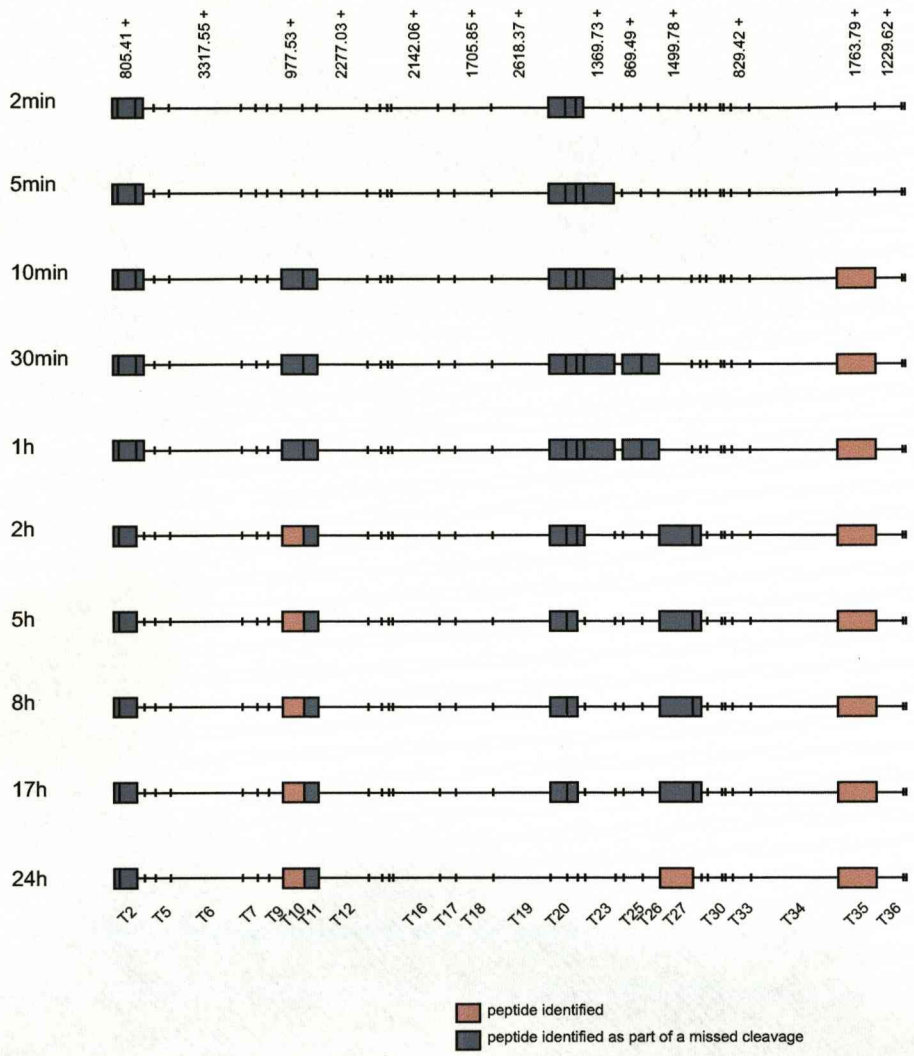



**Figure 94. Proteolysis of GAPDH with trypsin.**
Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, UK; 1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) was digested with trypsin (trypsin:protein 1:500) over 24h at 37°C. Proteolysis was stopped at 0, 2, 5, 10, 30, 60, 120, 300, 480, 1020 and 1440 min by mixing 10µL from the digestion mixture with 10µL 10% (v/v) formic acid. The resulting peptides were analysed by MALDI-ToF MS and peptides identified independently, or as part of a missed cleavage were used to construct peptide maps.

It followed therefore that prior denaturation of the protein might permit deamidation prior to digestion with trypsin. To confirm this, GAPDH was denatured by heating to 60°C for 1h before proteolysis, a denaturation treatment that was not sufficient to cause the protein to precipitate. Subsequently, when trypsin was added, the N-terminal peptide was again released rapidly, and the proportion of amide and acid variants of the peptide was assessed as previously described (Figure 95). Under these circumstances, the peptide first released was approximately 80% amide, with a significant proportion of acid form being measurable. This contrasted markedly with proteolysis of the native protein, when the peptide is initially all in the amide form. This behaviour most likely reflects the increased conformational flexibility of the peptide in the heat-treated protein, such that the peptide could acquire a conformation that allowed deamidation. Further, this unfolded and flexible component might be expected to be hypersensitive to proteolysis and to be released first. As the digestion proceeded, additional peptide in the amide form was released, and the proportion of amide therefore increased transiently, until the deamidation reaction dominated the peptide profile. Using the functions derived previously, a value for deamidation of $0.0023min^{-1}$ was obtained, in close agreement with that observed previously. If the heat-treated peptide was allowed to incubate at 37°C for 24h after the 60min denaturation period at 60°C, and then proteolysed with trypsin, the peptide first released was now only 50% in the amide form, consistent with extensive deamidation prior to proteolysis, consequential to denaturation. Again, as expected, proteolysis led to the slower release of peptide that was constrained and unable to deamidate and there was a transient increase in the proportion of amide which again decayed at the same rate as observed previously ($k_2$=$0.0024min^{-1}$). The behaviour of the system was consistent with the GAPDH preparation being 76% in the amide form, and 26% in a denatured form that was then rapidly proteolysed to generate the free acid form of the peptide. The effect of denaturation on the availability of the N-terminal peptide of GAPDH for deamidation is quite striking and defines the importance of monitoring the two processes of proteolysis and deamidation simultaneously, especially as this effect is only observed upon loss of higher order structure, and not upon increasing concentration of protease (results not shown).

The implications of the post-translational modification are several fold; for protein identification by peptide mass fingerprinting, deamidated peptides with a consequent change in isotope profile will not be matched to monoisotopic masses in the database for that particular protein, causing anomalies for protein identification. For characterisation proteomics, observation of a deamidation event may be incorrectly assigned as occurring *in vivo* when it is likely that

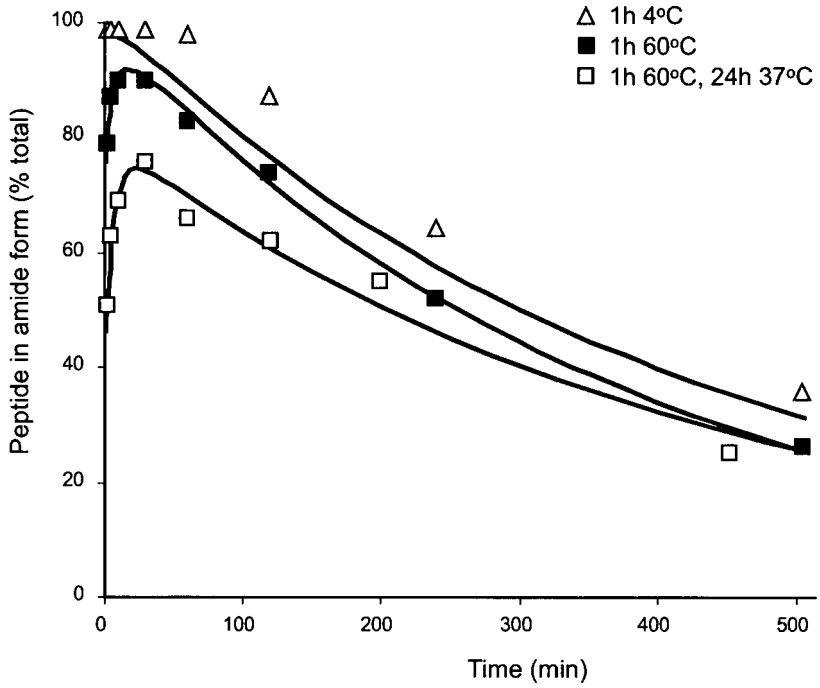



**Figure 95. The effect of denaturing protein structure by heating on the rate of deamidation.** GAPDH (1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) purified from rabbit skeletal muscle (Sigma, Dorset, UK) was digested with trypsin in-solution at a ratio trypsin:protein 1:100 at 37°C for 24h. Prior to digestion, GAPDH was incubated for 1h at 4°C (open triangles), 1h at 60°C (closed squares) and 1h at 60°C followed by 24h at 37°C (open squares). For each, deamidation was monitored over 24h of proteolysis and the proportion of acid and amide was calculated from the relative peak intensities of the two ions in MALDI-ToF mass spectra.

standard conditions for in-gel and in-solution proteolysis may promote deamidation, for which the constraining effect of three dimensional structure, limiting availability of potential sites for deamidation is apparent. For protein quantification using chemically synthesised internal standard peptides containing stable isotopes, or QconCAT proteins, an analyte or internal standard peptide seen to deamidate post-proteolysis may still be used by incorporating the signal for both acid and amide variants into calculations of relative signal intensity. For this, it is important that both asparagine and aspartic acid containing peptide are both ionised to the same extent. This information may also be incorporated into design of a QconCAT experiment, to avoid peptides containing asparagine, or to carry out preliminary experiments to look for deamidation in the analyte system. This investigation of deamidation of the GAPDH N-terminal peptide was complicated by the presence of a missed cleavage, and thus quantification based on this peptide as an internal standard is not reliable. To confirm the effects of this secondary proteolysis, a known amount of GAPDH purified from rabbit muscle was digested with trypsin. The stable isotope labelled internal standard peptide used for quantification of deamidation (Figure 96) was added prior to the addition of protease, and digestion was stopped at selected time points by mixing with 10% (v/v) formic acid. Peptides were analysed by MALDI-ToF MS and the entire peak area of the peptide envelopes for both analyte and internal standard was summed in each case for quantification based on relative peak intensity. This highlighted that, as expected, quantification became progressively worse with longer incubation times with trypsin although correlation was always strong, indicating that a correction factor could be applied based on the rate of proteolysis at the internal cleavage site, if necessary, although this would be undesirable if more suitable peptides are available for quantification.

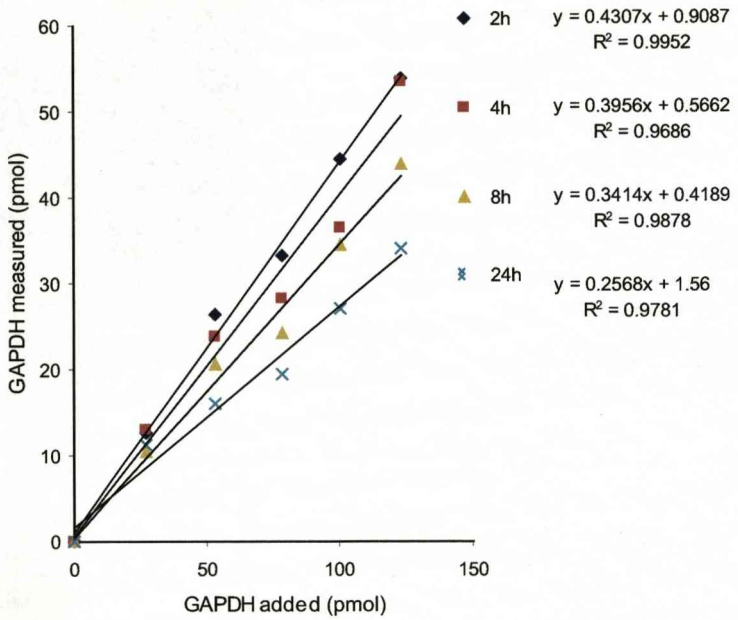



**Figure 96. Absolute quantification of GAPDH based on a deamidating peptide containing an internal trypsin cleavage site.**

Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, UK; 1mg/mL diluted to 0.2mg/mL with 50mM ammonium bicarbonate) was digested with trypsin (trypsin:protein 1:100) in increasing amounts from 0 to 54pmol over 24h at 37°C. The N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and mass 1041.59Da, was synthesised by Sigma-Genosys (Dorset, UK) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10Da mass offset from the analyte peptide and this was added to GAPDH in increasing amounts from 0 to 8pmol prior to addition of trypsin. Proteolysis was stopped at 2, 4, 8 and 24h by mixing 10µL from the digestion mixture with 10µL 10% (v/v) formic acid. The resulting peptides were analysed by MALDI-ToF MS and the entire peak area of the peptide envelopes for both analyte and internal standard was summed in each case for quantification based on relative peak intensity, thus incorporating both acid and amide variants.

# 6. Conclusions

## CONCLUSIONS

To develop proteomics as a quantitative science, several strategies have been employed. These include gel based methods, western blotting, mass tagging, mass spectrometry with prior peptide separation, and standardisation using stable isotope labelled reference molecules. To achieve quantification of proteins using the highly sensitive technique of mass spectrometry, proteins are most commonly digested into peptides with a specific protease, for example trypsin. Mass spectrometry of peptides is not inherently quantitative, thus stable isotopes are used to compensate for differences due to effects of ionisation, mass analysis and detection. Stable isotopes are incorporated into proteins and peptides by differential labelling, incorporation of a labelled 'tag', or by metabolic incorporation; for cells in culture or intact animals (reviewed by Bantscheff *et al.,* 2007). Strategies that avoid the use of stable isotopes for protein quantification are also under development, using either peptide signal intensity, or counting the number of fragment ion spectra used to identify a particular protein. For these methods, spectral counting is more sensitive for detecting changes in protein abundanc,e but calculations of peptide peak area intensity are more accurate for determining protein ratios (Old *et al.,* 2005).

To progress from relative quantification to absolute quantification in which proteins are expressed as for example, number of molecules per cell, or nmol per gram of tissue, internal standards that are either proteins or peptides with incorporated stable isotopes are mixed with analyte proteins. For the quantification of individual proteins, the true internal standard would be the corresponding protein expressed in both pure and stable isotope labelled form for quantification. Although both analyte and internal standard are subjected to the same sample processing, efficient expression of multiple native proteins would be challenging, as would subsequent purification and MS analysis of complex isotope envelopes. Alternatively, stable isotope labelled proteins are used as internal standards for quantification based on limit peptides, with the advantage that both analyte and internal standard peptides are present within the same sequence context, negating the need to take control of proteolysis separately for both analyte and standard proteins (Brun *et al.,* 2007). Alternatively, peptide-based approaches using proteotypic peptides as surrogates for the protein of interest have been developed. To create the appropriate internal standard, the selected surrogate peptide is chemically synthesised with incorporated stable isotopes. This is added to the analyte prior to mass spectrometric detection in a known amount and the relative signal intensity of the analyte

and internal standard peptide is used for absolute quantification of the native protein. This approach is well designed for the quantification of proteins in biological material that cannot be pre-labelled but for the quantification of multiple proteins, each would have to be independently quantified using a separately synthesised internal standard. To multiplex this approach, a peptide is selected from several proteins to be quantified and these sequences are concatenated into a synthetic gene which is expressed in a host organism (usually bacterial), labelled with stable isotopes, purified and digested with a protease to create a series of internal standards in equimolar amounts; the amount of each peptide is given from quantification of the recombinant **Q**con**CAT** (quantification concatamer) protein (Beynon *et al.,* 2005, Pratt *et al.,* 2006). For generation of multiple stable isotope labelled internal standard peptides, biological synthesis of recombinant proteins is preferred to chemical synthesis of peptides, the success of which is sequence dependent (for example hydrophobic peptides are difficult to synthesise chemically). Metabolic labelling is also advantageous for the incorporation of a variety of stable isotopes, with straightforward production of unlabelled and labelled QconCAT proteins for use in assessing assumptions of the method in addition to absolute quantification of proteins, or as a set of multiple internal standards for *in vitro* labelling.

There are two main components of a QconCAT method, a design phase and an implementation phase. The design phase predominantly involves nomination of proteins to be quantified, and selection of peptide sequences to incorporate into the QconCAT gene, for which there are a number of considerations. First is the protease to be used for peptide hydrolysis as this will determine the potential peptides for each protein to be quantified. Most proteomics experiments use trypsin as this is a highly specific enzyme cleaving at the C-terminus of lysine and arginine residues, generating the majority of peptides within the mass range of common instrumentation. The use of an alternative protease, for example Arg-C offers advantages, including greater probability of efficient ionisation and production of fewer peptides, reducing complexity but there may not be sufficient peptides within the desired mass range to select one that is unique for each protein to be quantified. Arg-C is also more expensive than trypsin for routine use, an important consideration for experimental design. Consequently, protease selection for the QconCAT method can only be determined from the specific requirements of the analyte system. For example, hydrophobic membrane proteins, especially those spanning the membrane may have limited accessibility to protease cleavage sites, in which case solubilisation and alternative cleavage strategies (reviewed by Wu and Yates, 2003) may be considered. Selection of Q-peptides is based on uniqueness of mass and

efficiency of ionisation, with appropriate sites for incorporation of stable isotope(s). As such, this can either be predicted on the basis of physicochemical characteristics with prior knowledge of the consequences of ionisation for certain amino acids, or from previous experience. Strategies for predicting proteotypic peptides are in early stages of development (Mallick *et al.,* 2007), although these look promising and would negate the need for preliminary investigation for peptide selection. However currently, prior knowledge of analyte peptides detected in the analytical environment, particularly using sample preparation protocols and instrumentation of choice to ensure that peptides are selected based on their propensity to ionise is necessary. Before incorporation into a QconCAT, selected peptides should be sequenced to confirm their identity and any possible modifications should be investigated. For example, peptides containing methionine are avoided where possible as variable oxidation of this residue would compromise quantification based on the signal intensity of the unmodified peptide. If it is not appropriate to conduct extensive investigation of potential proteotypic peptides prior to incorporation into a QconCAT gene, more than one Q-peptide could be selected per protein to improve the chance of reliable quantification using one or the other. It is also possible to enhance the ionisation of Q-peptides that are under-represented following MS analysis, for example by guanidination of lysine residues (for MALDI-ToF MS). Once the peptide sequences for incorporation into the QconCAT gene have been determined, it is necessary to optimise the codon sequence for efficient expression in the appropriate vector through the formation of stable RNA secondary structure. This is achieved on a trial and error basis, with random concatenation *in silico* and analysis of the predicted transcript for RNA secondary structure that might diminish expression; the order of peptides is altered if necessary. It is also advantageous to add sacrificial peptide sequences to the N- and C-terminus of the QconCAT gene to protect the true quantification peptides from exoproteolytic attack during expression. This also provides an opportunity to add an initiator methionine residue to the N-terminus, and other desirable features, for example a purification HisTag. Henceforth, expression in the vector system (for example *E.coli*) should be fairly straightforward with no higher order folding and this provides the ideal opportunity for incorporation of a stable isotope label with supplementation into the medium for *E.coli* growth. The labelling strategy must also be appropriately designed, and the decision must be made whether to employ a uniform label to ensure that every peptide is labelled with a different mass offset from the analyte peptide (for example [$^{15}$N]), or to label specific amino acids, preferably in every peptide (for example [$^{13}C_6$]lys/arg for tryptic peptides). It is important that the mass offset between analyte and internal standard can be distinguished by mass spectrometry and

consequently it is advisable that this is greater than 4Da. It is also possible to label QconCAT proteins using *in vitro* labelling strategies if necessary. For the strategy chosen, efficiency of labelling can be confirmed from the peptide isotopomer envelope in MS analysis. Once designed, the QconCAT method is implemented by expression, labelling and purification prior to use as a set of internal standards for absolute quantification.

There are many challenges for absolute quantification, using either the synthetic peptide, or QconCAT approach. First, is the selection of the peptides that are to be used as surrogate standards for the analyte proteins. In the future this may be possible to predict without prior experimental work, based on physicochemical characteristics of peptides and their known behaviour in different analytical environments. Currently this is based on previous experience of detected peptides, and those that are unique, within a given mass range and ionise most efficiently are often chosen. For an unknown analytical system, this could be achieved using a one off shotgun proteomics experiment from which a list of signature peptides can be identified (for example, Wienkoop and Weckwerth, 2006).

A second challenge for these strategies is the attainment of complete proteolysis of analyte and QconCAT proteins, vital for protein quantification based on the amount of a single limit peptide. Peptides incorporated into a QconCAT exist in a different sequence context to their native environment, and consequently the rate of hydrolysis will be different for analyte and internal standard proteins. As QconCAT proteins generally lack higher order structure following recombinant expression, they are rapidly digested, independent of peptide order in the QconCAT (Mirzaei *et al.,* 2007). However analyte protein digestion is hindered by higher order structure which restricts access to enzymatic cleavage sites, resulting in much slower rates of proteolysis (Rivers *et al.,* 2007). For absolute quantification using QconCAT proteins or synthetic peptides, it is crucial that complete digestion is achieved. If this is not observed for native proteins, analyte proteins must be sufficiently denatured, or conditions, for example enzyme concentration, incubation time and presence of organic solvents, must be altered to achieve complete proteolysis. The rate of proteolysis of analyte proteins can be derived using pre-digested QconCAT labelled peptides, or stable isotope labelled synthetic peptides as internal standards, and this is highly recommended to ensure complete digestion is achieved for reliable quantification.

Thirdly, is the challenge of complex biological sample dynamic range of protein abundance, which for human plasma covers ten orders of magnitude (Anderson and Anderson, 2002). Consequently, only a subset of proteins are analysed, and of those, only a subset of peptides are quantifiable. For quantitative proteomics, current strategies have not achieved quantification of an entire proteome, and are usually targeted to high abundance proteins. Low abundance proteins often provide valuable information, particularly in the discovery of clinically relevant biomarkers of disease and these are difficult to detect due to the domination of high abundance species in all forms of analysis. Methods to enrich low abundance proteins, or deplete proteins in high abundance are effective, but this usually loses quantitative information as protein abundance no longer reflects natural expression. In a single experiment, dynamic range is limited by capacity and resolution of instrumentation, but also by the abundance range that can be quantified using a single internal standard. For quantification using QconCAT in a single experiment, dynamic range of abundance covering over two orders of magnitude (0-300nmol/g tissue) is quantifiable (Rivers *et al.,* 2007). For quantification of proteins beyond this range in abundance, different amounts of QconCAT protein may be added to the analyte sample, quantifying low and high abundance proteins in separate experiments. This will also be improved by chromatographic separation of peptides prior to MS analysis to remove suppression of higher abundance peptides. In this instance, quantification using stable isotope labelled synthetic peptides may be preferred as a different amount of each peptide can be added for the quantification of each protein, depending on the natural abundance of the analyte.

Another significant challenge for quantitative proteomics is quantification of post-translational modifications (reviewed by Chen *et al.,* 2007). These are often important for regulating cell signalling events and protein interactions and are difficult to monitor due to their dynamic and often reversible nature, for example phosphorylation. Strategies to quantify phosphorylation using stable isotope labelling include incorporation of selectively labelled amino acids that are common targets for phosphorylation in cell culture, for example [$^{2}H_3$]serine (Zhu *et al.,* 2002), or differentially labelling cell populations phosphorylated to different extents with combinations of [$^{13}C_6$] and [$^{15}N_4$] arginine prior to affinity purification of phosphopeptides (Blagoev *et al.,* 2004). Absolute quantification has been achieved using stable isotope labelled synthetic peptides (Gerber *et al.,* 2003), but this was achieved using a synthetic peptide containing the specific modification. Such analysis is limited with the QconCAT strategy because it is designed to quantify the unmodified analyte peptide. For standard quantification experiments, peptides

known to be modified post-translation are avoided for incorporation into the QconCAT as a partial modification would sequester a proportion of the peptide signal, thus compromising quantification. To overcome this for well-characterised, single sites of modification, a peptide may be incorporated into the QconCAT in both native and modifiable states. This would permit quantification of the amount of modification by comparing relative signal intensities of the two analyte peptides with the relevant internal standard. Alternatively, the modified peptide sequence could be incorporated into the QconCAT with the analyte peptide quantified before and after treatment to remove the modification, for example alkaline phosphatase to hydrolyse phosphopeptides.

For quantification of multiple proteins in a complex biological system, the QconCAT approach is preferable to the use of several stable isotope labelled synthetic peptides. Many of the advantages of this approach have been discussed, with more reliability achieved by mixing internal standards with analyte samples at the protein level, controlling effects of sample preparation as early as possible. Synthetic peptides are often incorporated after proteolytic digestion of analyte; they could also be added at the protein level but they are likely to bind to tube walls, become degraded or modified, especially with long incubation times necessary for proteolysis (Mirzaei *et al.*, 2007). From calculations of the relative cost of the two methods, the synthetic peptide approach is more economical for quantification of less than 10 proteins, but for an average QconCAT protein containing 50 peptides, the QconCAT approach is approximately 15% of the cost of chemical synthesis of each peptide in stable isotope labelled form. Futhermore, a single preparation of QconCAT yields approximately 250nmol of protein (and hence Q-peptides), compared to 5x1nmol typically supplied by the manufacturer of synthetic peptides. For the QconCAT method, initial expense of the construction of the QconCAT gene is offset by the repeated use of this construct for protein expression, for example to test different labelling strategies and produce multiple preparations of the recombinant protein. For reliability of quantification using these two strategies, quantification has been compared (Rivers *et al.*, 2007, Mirzaei *et al.*, 2007) with similar results achieved for both methods. Small discrepancies are attributed to differences in the way that the two internal standards are quantified. Quantification of synthetic peptides is achieved using amino acid analysis, conducted by the manufacturer. This uses highly sensitive techniques but it is difficult to ensure all supplied peptide is recovered into solution and the low quantity of peptide provided is not sufficient to quantify by the end user. There are a number of methods for quantification of the QconCAT protein, some of which are more sensitive but require sacrifice of

a greater proportion of protein product. For most analyses, quantification by protein assay is sufficient relative to the analyte system, for example compared to biological variance from absolute quantification of chicken skeletal muscle soluble proteins, coefficient of variation for analytical replicates was very low, and experiments to assess accuracy confirmed the reliability of quantification (Rivers *et al.,* 2007). For application of the QconCAT method across multiple platforms, quantifying a range of biological proteins, quantification of the internal standard protein must be highly accurate. This could be achieved using a synthetic peptide purchased separately (unlabelled) for the quantification of a specific peptide that is incorporated into every QconCAT protein. If every QconCAT protein is quantified by the synthetic peptide in MS, quantification will be reliable across different laboratories covering a host of different applications.

Successful applications of absolute quantification using the QconCAT method are widespread, from the analysis of protein expression of different isoforms with single amino acid differences, to the assessment of defined peptides in MS and MSMS for instrument validation or assessment of ionisation effects. The number of peptides that can be incorporated into a single QconCAT is restricted as high-level expression of large proteins is more problematical. However, QconCAT proteins are currently in production (unpublished data) containing in the region of 100 Q-peptides in the mass range 800-4000Da producing a QconCAT protein of approximately 150kDa. For optimal expression, it may be preferable to create two QconCAT genes of smaller size but this is under investigation.

The objective of this research was to provide a rigorous test of the QconCAT method for absolute quantification of multiple proteins, incorporating an in-depth assessment of variance within this system. This method was applied successfully for the quantification of multiple proteins in the soluble fraction of chicken skeletal muscle from 0 to 30d growth in both broiler and layer strains. Protein abundance was assessed at six time points during growth, with four birds at each time point. The QconCAT is robust to the choice of mass spectrometric platform used with detection of some proteins only by one instrument, with or without prior chromatographic separation. Accuracy of the method was also assessed, in addition to variation contributed by biological and analytical replicate experiments. Analytical variance was significantly lower than that for quantification data measured for four different birds of each strain, for example the analytical variance (CV of 6.0% for β-enolase, n=4) compared favourably to biological variance (CV of 24.0% for β-enolase, n=4). Perhaps the most

significant limitation of the QconCAT method, is the dynamic range that can be achieved in a single experiment, which for absolute quantification of chicken skeletal muscle soluble proteins covered two orders of magnitude of protein abundance. This has been discussed in more detail previously, and compared to the synthetic peptide method for absolute quantification. Through the application of QconCAT to this biological system, some advantages of absolute quantification using stable isotope labelled surrogate peptides as internal standards as a general tool for quantitative proteomics have emerged. Firstly to quantify the normalisation of protein abundance using ligand library beads, and secondly to provide a more detailed insight into the post-translational modification, deamidation, ascertaining the relationship between higher order protein structure and deamidation (Rivers *et al.,* 2008).

Quantitative strategies, particularly for absolute quantification, and their application have been discussed (recently reviewed by Bantscheff *et al.,* 2007). The greatest challenge for these approaches, is to achieve the comprehensive quantification of a biological system and future work in this area should be directed to achieving this goal. For a QconCAT approach, a relatively simple proteome, for example *E.coli* containing in the region of 4-5,000 proteins (strain K12, Swissprot database; www.expasy.org/sprot), with one peptide per protein would require 4,500 Q-peptides for quantification. If 100 peptides can be incorporated into a single QconCAT protein, this would require 45 QconCAT proteins, each of approximately 100-150kDa. This is a relatively small number of gene design and synthesis procedures, relative to the chemical synthesis of 4,500 peptides required for the alternative strategy to achieve absolute quantification using MS. To achieve this, the QconCAT strategy is the most efficient, at 15% of the cost (based on in-house expression, labelling and purification). In addition, once the QconCAT genes have been designed, this provides an unlimited resource of multiple internal standard peptides that can be expressed with a variety of incorporated stable isotopes. For absolute quantification, QconCAT genes can be designed for specific functional groups of proteins, for example membrane bound proteins, transporter proteins, or specific enzymes. In addition, different QconCAT experiments can be implemented for proteins varying in abundance, with appropriately designed labelling strategies for each. As such, this methodology has the potential to become the 'gold-standard' for global absolute quantification, although this technology will still encounter certain challenges. The achievement of quantification of an entire proteome using the QconCAT strategy must be in accordance with developments in the ability to predict proteotypic peptides for selection and incorporation into the QconCAT, a significant advantage for this type of approach, negating the need for

extensive prior investigation of the analyte system. However, this may require the incorporation of two Q-peptides per protein to be quantified, thus doubling the number of QconCAT proteins needed for absolute quantification. Furthermore, detection and quantification of 4,500 (or 9,000 if two Q-peptides per protein are incorporated) peptide pairs requires significant advances in statistical treatment and interpretation of quantitative data, such that this approach can become automated, and used for high throughput quantification. For large scale programmes, and the implementation of the QconCAT strategy for a wide variety of applications across multiple platforms, quantification of internal standard proteins, for example incorporation of a universal synthetic peptide is essential. As a further obstacle for system-wide quantification using QconCAT, sample complexity and dynamic range poses a substantial challenge. To delve further into the proteome, MS methods for multiple reaction monitoring using triple quadrupole instruments where the intact peptide mass and one or more specific fragment ions in an LC-MS experiment are monitored have been developed. This technique essentially uses two mass filters to increase specificity and sensitivity for absolute quantification using stable isotope labelled reference peptides. This permits a greater dynamic range of protein abundance to be detected and quantified, for high (55mg/mL) and medium (1µg/mL) abundance peptides covering 4-5 orders of magnitude (Anderson and Hunter, 2006) and as low as 1-10ng/mL without antibody enrichment (Keshishian *et al.,* 2007). This strategy is becoming widely used for routine protein quantification, with particular application to clinical diagnostics (for example Kirsch *et al.,* 2007) and its application to quantification using QconCAT proteins has the potential to improve the dynamic range of protein abundance quantified in a single experiment considerably.

For understanding of complex biological systems to drive the field of 'systems biology', absolute quantification of proteins is essential. With continued development of the QconCAT method for various applications and large-scale proteomic analysis, absolute quantification of the entire proteome of an organism is within reach. To realise this goal, particular focus should be directed to strategies for predicting proteotypic peptides, computational tools to handle substantial amounts of quantitative data, and advances in instrumentation and strategies to quantify proteins covering a wide dynamic range of abundance.

# 7. REFERENCES

## 7. REFERENCES


Adams, J.M. and Capecchi, M.R. (**1966**). "N-formylmethionyl-sRNA as the initiator of protein synthesis." *Proc Natl Acad Sci U S A* 55(1): 147-55.

Adams, J.M. (**1968**). "On the release of the formyl group from nascent protein." *J Mol Biol* 33(3): 571-89.

Aebersold, R. (**2003**). "Constellations in a cellular universe." *Nature* 422(6928): 115-6.

Anderson, N.L. and Anderson, N.G. (**2002**). "The human plasma proteome: history, character, and diagnostic prospects." *Mol Cell Proteomics* 1(11): 845-67.

Anderson, N.L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R.S., Conrads, T.P., Veenstra, T.D., Adkins, J.N., Pounds, J.G., Fagan, R. and Lobley, A. (**2004**). "The human plasma proteome: a nonredundant list developed by combination of four separate sources." *Mol Cell Proteomics* 3(4): 311-26.

Anderson, L. and Hunter, C.L. (**2006**). "Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins." *Mol Cell Proteomics* 5(4): 573-88.

Balogh, M.P. (**2004**). "Debating resolution and mass accuracy in mass spectrometry." *Spectroscopy* 19(10): 34-40.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. and Kuster, B. (**2007**). "Quantitative mass spectrometry in proteomics: a critical review." *Anal Bioanal Chem* 389(4): 1017-31.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y. and Barkai, N. (**2006**). "Noise in protein expression scales with natural protein abundance." *Nat Genet* 38(6): 636-43.

Barr, J.R., Maggio, V.L., Patterson, D.G., Jr., Cooper, G.R., Henderson, L.O., Turner, W.E., Smith, S.J., Hannon, W.H., Needham, L.L. and Sampson, E.J. (**1996**). "Isotope dilution-mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I." *Clin Chem* 42(10): 1676-82.

Baumgart, S., Lindner, Y., Kuhne, R., Oberemm, A., Wenschuh, H. and Krause, E. (**2004**). "The contributions of specific amino acid side chains to signal intensities of peptides in matrix-assisted laser desorption/ionization mass spectrometry." *Rapid Commun Mass Spectrom* 18(8): 863-8.

Belle, A., Tanay, A., Bitincka, L., Shamir, R. and O'Shea, E.K. (**2006**). "Quantification of protein half-lives in the budding yeast proteome." *Proc Natl Acad Sci U S A* 103(35): 13004-9.

Ben-Bassat, A., Bauer, K., Chang, S.Y., Myambo, K., Boosman, A. and Chang, S. (**1987**). "Processing of the initiation methionine from proteins: properties of the *Escherichia coli* methionine aminopeptidase and its gene structure." *J Bacteriol* 169(2): 751-7.

Beynon, R.J. (**2005**). "A simple tool for drawing proteolytic peptide maps." *Bioinformatics* 21(5): 674-5.

Beynon, R.J., Doherty, M.K., Pratt, J.M. and Gaskell, S.J. (**2005**). "Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides." *Nat Methods* 2(8): 587-9.

Beynon, R.J. and Pratt, J.M. (**2005**). "Metabolic labeling of proteins for proteomics." *Mol Cell Proteomics* 4(7): 857-72.

Biringer, R.G., Amato, H., Harrington, M.G., Fonteh, A.N., Riggins, J.N. and Huhmer, A.F. (**2006**). "Enhanced sequence coverage of proteins in human cerebrospinal fluid using multiple enzymatic digestion and linear ion trap LC-MS/MS." *Brief Funct Genomic Proteomic* 5(2): 144-53.

Blagoev, B., Ong, S.E., Kratchmarova, I. and Mann, M. (**2004**). "Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics." *Nat Biotechnol* 22(9): 1139-45.

Bouwman, F., Renes, J. and Mariman, E. (**2004**). "A combination of protein profiling and isotopomer analysis using matrix-assisted laser desorption/ionization-time of flight mass spectrometry reveals an active metabolism of the extracellular matrix of 3T3-L1 adipocytes." *Proteomics* 4(12): 3855-63.

Browne, T.R. (**1986**). "Stable isotopes in pharmacology studies: present and future." *J Clin Pharmacol* 26(6): 485-9.

Brun, V., Dupuis, A., Adrait, A., Marcellin, M., Thomas, D., Court, M., Vandenesch, F. and Garin, J. (**2007**). "Isotope-labeled protein standards: toward absolute quantitative proteomics." *Mol Cell Proteomics* 6(12): 2139-49.

Byrne, A.R., Benedick, L. (**1997**). "An internal standard method in alpha spectrometric determination of uranium and thorium radioisotopes using instrumental neutron activation analysis." *Analytical Chemistry* 69(6): 996-999.

Cagney, G. and Emili, A. (**2002**). "De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging." *Nat Biotechnol* 20(2): 163-70.

Cagney, G., Amiri, S., Premawaradena, T., Lindo, M. and Emili, A. (**2003**). "In silico proteome analysis to facilitate proteomics experiments using mass spectrometry." *Proteome Sci* 1(1): 5.

Canas, B., Lopez-Ferrer, D., Ramos-Fernandez, A., Camafeita, E. and Calvo, E. (**2006**). "Mass spectrometry technologies for proteomics." *Brief Funct Genomic Proteomic* 4(4): 295-320.

Castagna, A., Cecconi, D., Sennels, L., Rappsilber, J., Guerrier, L., Fortis, F., Boschetti, E., Lomas, L. and Righetti, P.G. (**2005**). "Exploring the hidden human urinary proteome via ligand library beads." *J Proteome Res* 4(6): 1917-30.

Chait, B.T. (**2006**). "Chemistry. Mass spectrometry: bottom-up or top-down?" *Science* 314(5796): 65-6.

Chamrad, D. and Meyer, H.E. (**2005**). "Valid data from large-scale proteomics studies." *Nat Methods* 2(9): 647-8.

Chen, X., Sun, L., Yu, Y., Xue, Y. and Yang, P. (**2007**). "Amino acid-coded tagging approaches in quantitative proteomics." *Expert Rev Proteomics* 4(1): 25-37.

Chong, P.K., Gan, C.S., Pham, T.K. and Wright, P.C. (**2006**). "Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections." *J Proteome Res* 5(5): 1232-40.

Claverol, S., Burlet-Schiltz, O., Gairin, J.E. and Monsarrat, B. (**2003**). "Characterization of protein variants and post-translational modifications: ESI-$MS^n$ analyses of intact proteins eluted from polyacrylamide gels." *Mol Cell Proteomics* 2(8): 483-93.

Cox, J. and Mann, M. (**2007**). "Is proteomics the new genomics?" *Cell* 130(3): 395-8.

Craig, R. and Beavis, R.C. (**2004**). "TANDEM: matching proteins with tandem mass spectra." *Bioinformatics* 20(9): 1466-7.

de Hoog, C.L. and Mann, M. (**2004**). "Proteomics." *Annu Rev Genomics Hum Genet* 5: 267-93.

Deverman, B.E., Cook, B.L., Manson, S.R., Niederhoff, R.A., Langer, E.M., Rosova, I., Kulans, L.A., Fu, X., Weinberg, J.S., Heinecke, J.W., Roth, K.A. and Weintraub, S.J. (**2002**). "Bcl-$x_L$ deamidation is a critical switch in the regulation of the response to DNA damage." *Cell* 111(1): 51-62.

Doherty, M.K., Mclean, L., Hayter, J.R., Pratt, J.M., Robertson, D.H., El-Shafei, A., Gaskell, S.J. and Beynon, R.J. (**2004**). "The proteome of chicken skeletal muscle: changes in soluble protein expression during growth in a layer strain." *Proteomics* 4(7): 2082-93.

Doherty, M.K., Whitehead, C., Mccormack, H., Gaskell, S.J. and Beynon, R.J. (**2005**). "Proteome dynamics in complex organisms: using stable isotopes to monitor individual protein turnover rates." *Proteomics* 5(2): 522-33.

Elias, J.E., Haas, W., Faherty, B.K. and Gygi, S.P. (**2005**). "Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations." *Nat Methods* 2(9): 667-75.

Ellison, D., Hinton, J., Hubbard, S.J. and Beynon, R.J. (**1995**). "Limited proteolysis of native proteins: the interaction between avidin and proteinase K." *Protein Sci* 4(7): 1337-45.

Eng, J.K.M., A. L. Yates, J. R. (**1994**). "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." *J Am Soc Mass Spectrom* 5: 976-989.

Fanara, P., Turner, S., Busch, R., Killion, S., Awada, M., Turner, H., Mahsut, A., Laprade, K.L., Stark, J.M. and Hellerstein, M.K. (**2004**). "*In vivo* measurement of microtubule dynamics using stable isotope labeling with heavy water. Effect of taxanes." *J Biol Chem* 279(48): 49940-7.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. and Whitehouse, C.M. (**1989**). "Electrospray ionization for mass spectrometry of large biomolecules." *Science* 246(4926): 64-71.

Flannery, A.V., Easterby, J.S. and Beynon, R.J. (**1992**). "Turnover of glycogen phosphorylase in the pectoralis muscle of broiler and layer chickens." *Biochem J* 286 ( Pt 3): 915-22.

Friedman, A.R., Ichhpurani, A.K., Brown, D.M., Hillman, R.M., Krabill, L.F., Martin, R.A., Zurcher-Neely, H.A. and Guido, D.M. (**1991**). "Degradation of growth hormone releasing factor analogs in neutral aqueous solution is related to deamidation of asparagine residues. Replacement of asparagine residues by serine stabilizes." *Int J Pept Protein Res* 37(1): 14-20.

Galperin, M.Y. and Koonin, E.V. (**2004**). "'Conserved hypothetical' proteins: prioritization of targets for experimental study." *Nucleic Acids Res* 32(18): 5452-63.

Gan, C.S., Chong, P.K., Pham, T.K. and Wright, P.C. (**2007**). "Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ)." *J Proteome Res* 6(2): 821-7.

Gannes, L.Z., Martinez Del Rio, C. and Koch, P. (**1998**). "Natural abundance variations in stable isotopes and their potential uses in animal physiological ecology." *Comp Biochem Physiol A Mol Integr Physiol* 119(3): 725-37.

Geiger, T. and Clarke, S. (**1987**). "Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation." *J Biol Chem* 262(2): 785-94.

Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. and Gygi, S.P. (**2003**). "Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS." *Proc Natl Acad Sci U S A* 100(12): 6940-5.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (**2003**). "Global analysis of protein expression in yeast." *Nature* 425(6959): 737-41.

Goodlett, D.R., Keller, A., Watts, J.D., Newitt, R., Yi, E.C., Purvine, S., Eng, J.K., Von Haller, P., Aebersold, R. and Kolker, E. (**2001**). "Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation." *Rapid Commun Mass Spectrom* 15(14): 1214-21.

Graumann, J., Hubner, N.C., Kim, J.B., Ko, K., Moser, M., Kumar, C., Cox, J., Schoeler, H. and Mann, M. (**2007**). "SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins." *Mol Cell Proteomics.*

Griffin, H.D. and Goddard, C. (**1994**). "Rapidly growing broiler (meat-type) chickens: their origin and use for comparative studies of the regulation of growth." *Int J Biochem* 26(1): 19-28.

Guerrier, L., Thulasiraman, V., Castagna, A., Fortis, F., Lin, S., Lomas, L., Righetti, P.G. and Boschetti, E. (**2006**). "Reducing protein concentration range of biological samples using solid-phase ligand libraries." *J Chromatogr B Analyt Technol Biomed Life Sci* 833(1): 33-40.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. (**1999**). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." *Nat Biotechnol* 17(10): 994-9.

Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (**1999**). "Correlation between protein and mRNA abundance in yeast." *Mol Cell Biol* 19(3): 1720-30.

Gygi, S.P., Rist, B., Griffin, T.J., Eng, J. and Aebersold, R. (**2002**). "Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags." *J Proteome Res* 1(1): 47-54.

Hale, J.E., Butler, J.P., Knierman, M.D. and Becker, G.W. (**2000**). "Increased sensitivity of tryptic peptide detection by MALDI-TOF mass spectrometry is achieved by conversion of lysine to homoarginine." *Anal Biochem* 287(1): 110-7.

Halsey, J.F. and Harrington, W.F. (**1973**). "Substructure of paramyosin. Correlation of helix stability, trypsin digestion kinetics, and amino acid composition." *Biochemistry* 12(4): 693-701.

Harris, J.I. (**1956**). "Effect of urea on trypsin and alpha-chymotrypsin." *Nature* 177(4506): 471-3.

Hayter, J.R., Robertson, D.H.L., Gaskell, S.J. and Beynon, R.J. (**2003**). "Proteome analysis of intact proteins in complex mixtures." *Molecular & Cellular Proteomics* 2(2): 85-95.

Heller, M., Ye, M., Michel, P.E., Morier, P., Stalder, D., Junger, M.A., Aebersold, R., Reymond, F. and Rossier, J.S. (**2005**). "Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides." *J Proteome Res* 4(6): 2273-82.

Hellerstein, M.K. (**2004**). "New stable isotope-mass spectrometric techniques for measuring fluxes through intact metabolic pathways in mammalian systems: introduction of moving pictures into functional genomics and biochemical phenotyping." *Metab Eng* 6(1): 85-100.

Higgs, R.E., Knierman, M.D., Gelfanova, V., Butler, J.P. and Hale, J.E. (**2005**). "Comprehensive label-free method for the relative quantification of proteins from biological samples." *J Proteome Res* 4(4): 1442-50.

Hill, R.L., Schwartz, H.C. and Smith, E.L. (**1959**). "The effect of urea and guanidine hydrochloride on activity and optical rotation of crystalline papain." *J Biol Chem* 234(3): 572-6.

Hillenkamp, F., Karas, M., Beavis, R.C. and Chait, B.T. (**1991**). "Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers." *Anal Chem* 63(24): 1193A-1203A.

Hipple, J.A., Sommer, H., Thomas, H. A. (**1949**). "A precise method of determining the faraday by magnetic resonance." *Physical Review* 76: 1877.

Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M. and Graham Cooks, R. (**2005**). "The Orbitrap: a new mass spectrometer." *J Mass Spectrom* 40(4): 430-43.

Hu, L., Ye, M., Jiang, X., Feng, S. and Zou, H. (**2007**). "Advances in hyphenated analytical techniques for shotgun proteome and peptidome analysis-a review." *Anal Chim Acta* 598(2): 193-204.

Hubbard, S.J., Campbell, S.F. and Thornton, J.M. (**1991**). "Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors." *J Mol Biol* 220(2): 507-30.

Hubbard, S.J. (**1998**). "The structural aspects of limited proteolysis of native proteins." *Biochim Biophys Acta* 1382(2): 191-206.

Hubbard, S.J., Beynon, R.J. and Thornton, J.M. (**1998**). "Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures." *Protein Eng* 11(5): 349-59.

International Human Genome Consortium (**2004**). "Finishing the euchromatic sequence of the human genome." *Nature* 431(7011): 931-45.

Ishihama, Y., Sato, T., Tabata, T., Miyamoto, N., Sagane, K., Nagasu, T. and Oda, Y. (**2005**). "Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards." *Nat Biotechnol* 23(5): 617-21.

Jensen, P.K., Pasa-Tolic, L., Anderson, G.A., Horner, J.A., Lipton, M.S., Bruce, J.E. and Smith, R.D. (**1999**). "Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry." *Anal Chem* 71(11): 2076-84.

Ji, J., Chakraborty, A., Geng, M., Zhang, X., Amini, A., Bina, M. and Regnier, F. (**2000**). "Strategy for qualitative and quantitative analysis in proteomics based on signature peptides." *J Chromatogr B Biomed Sci Appl* 745(1): 197-210.

Johnson, R.S., Martin, S.A., Biemann, K., Stults, J.T. and Watson, J.T. (**1987**). "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine." *Anal Chem* 59(21): 2621-5.

Julka, S. and Regnier, F. (**2004**). "Quantification in proteomics through stable isotope coding: a review." *Journal of Proteome Research* 3(3): 350-63.

Keshishian, H., Addona, T., Burgess, M., Kuhn, E. and Carr, S.A. (**2007**). "Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution." *Mol Cell Proteomics* 6(12): 2212-29.

Kim, S.C., Chen, Y., Mirza, S., Xu, Y., Lee, J., Liu, P. and Zhao, Y. (**2006**). "A clean, more efficient method for in-solution digestion of protein mixtures without detergent or urea." *J Proteome Res* 5(12): 3446-52.

Kimmel, J.R. (**1967**). "Guanidination of proteins." *methods in enzymology* 11: 584-589.

Kirkpatrick, D.S., Gerber, S.A. and Gygi, S.P. (**2005**). "The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications." *Methods* 35(3): 265-73.

Kirsch, S., Widart, J., Louette, J., Focant, J.F. and De Pauw, E. (**2007**). "Development of an absolute quantification method targeting growth hormone biomarkers using liquid chromatography coupled to isotope dilution mass spectrometry." *J Chromatogr A* 1153(1-2): 300-6.

Kito, K., Ota, K., Fujita, T. and Ito, T. (**2007**). "A synthetic protein approach toward accurate mass spectrometric quantification of component stoichiometry of multiprotein complexes." *J Proteome Res* 6(2): 792-800.

Klammer, A.A. and MacCoss, M.J. (**2006**). "Effects of modified digestion schemes on the identification of proteins from complex mixtures." *J Proteome Res* 5(3): 695-700.

Kratzer, R., Eckerskorn, C., Karas, M. and Lottspeich, F. (**1998**). "Suppression effects in enzymatic peptide ladder sequencing using ultraviolet - matrix assisted laser desorption/ionization - mass spectormetry." *Electrophoresis* 19(11): 1910-9.

Krause, E., Wenschuh, H. and Jungblut, P.R. (**1999**). "The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins." *Anal Chem* 71(19): 4160-5.

Krijgsveld, J., Ketting, R.F., Mahmoudi, T., Johansen, J., Artal-Sanz, M., Verrijzer, C.P., Plasterk, R.H. and Heck, A.J. (**2003**). "Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics." *Nat Biotechnol* 21(8): 927-31.

Krokhin, O.V., Craig, R., Spicer, V., Ens, W., Standing, K.G., Beavis, R.C. and Wilkins, J.A. (**2004**). "An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS." *Mol Cell Proteomics* 3(9): 908-19.

Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W. and Guild, B. (**2004**). "Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and $^{13}C$-labeled peptide standards." *Proteomics* 4(4): 1175-86.

Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M., Pain, A., Sauerwein, R.W., Eling, W.M., Hall, N., Waters, A.P., Stunnenberg, H.G. and Mann, M. (**2002**). "Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry." *Nature* 419(6906): 537-42.

Li, J., Steen, H. and Gygi, S.P. (**2003**). "Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response." *Mol Cell Proteomics* 2(11): 1198-204.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R., 3rd (**1999**). "Direct analysis of protein complexes using mass spectrometry." *Nat Biotechnol* 17(7): 676-82.

Liu, P. and Regnier, F.E. (**2002**). "An isotope coding strategy for proteomics involving both amine and carboxyl group labeling." *J Proteome Res* 1(5): 443-50.

Liu, H., Sadygov, R.G. and Yates, J.R., 3rd (**2004**). "A model for random sampling and estimation of relative protein abundance in shotgun proteomics." *Anal Chem* 76(14): 4193-201.

Lopez-Ferrer, D., Ramos-Fernandez, A., Martinez-Bartolome, S., Garcia-Ruiz, P. and Vazquez, J. (**2006**). "Quantitative proteomics using $(^{16})O/(^{18})O$ labeling and linear ion trap mass spectrometry." *Proteomics* 6 Suppl 1: S4-S11.

Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E.M. (**2007**). "Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation." *Nat Biotechnol* 25(1): 117-24.

Lubec, G. and Afjehi-Sadat, L. (**2007**). "Limitations and pitfalls in protein identification by mass spectrometry." *Chem Rev* 107(8): 3568-84.

Luftig, R.B., Mcmillan, P.N. and Gudger, M. (**1974**). "Quantitation of endogenous C-type virion production in several murine cell lines." *J Virol* 14(4): 1017-21.

Lundell, N. and Schreitmuller, T. (**1999**). "Sample preparation for peptide mapping-A pharmaceutical quality-control perspective." *Anal Biochem* 266(1): 31-47.

Mallick, P., Schirle, M., Chen, S.S., Flory, M.R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B. and Aebersold, R. (**2007**). "Computational prediction of proteotypic peptides for quantitative proteomics." *Nat Biotechnol* 25(1): 125-31.

Mann, M. and Wilm, M. (**1994**). "Error-tolerant identification of peptides in sequence databases by peptide sequence tags." *Anal Chem* 66(24): 4390-9.

Mann, M., Hendrickson, R.C. and Pandey, A. (**2001**). "Analysis of proteins and proteomes by mass spectrometry." *Annu Rev Biochem* 70: 437-73.

March, R.E. (**2000**). "Quadrupole ion trap mass spectrometry: a view at the turn of the century." *Int J Mass Spec* 200: 285-312.

Marshall, A.G., Hendrickson, C.L. and Jackson, G.S. (**1998**). "Fourier transform ion cyclotron resonance mass spectrometry: a primer." *Mass Spectrom Rev* 17(1): 1-35.

McLaughlin, T., Siepen, J.A., Selley, J., Lynch, J.A., Lau, K.W., Yin, H., Gaskell, S.J. and Hubbard, S.J. (**2006**). "PepSeeker: a database of proteome peptide identifications for investigating fragmentation patterns." *Nucleic Acids Res* 34(Database issue): D649-54.

McLean, L., Doherty, M.K., Deeming, D.C. and Beynon, R.J. (**2004**). "A proteome analysis of the subcutaneous gel in avian hatchlings." *Mol Cell Proteomics* 3(3): 250-6.

McLean, L., Young, I.S., Doherty, M.K., Robertson, D.H., Cossins, A.R., Gracey, A.Y., Beynon, R.J. and Whitfield, P.D. (**2007**). "Global cooling: cold acclimation and the expression of soluble proteins in carp skeletal muscle." *Proteomics* 7(15): 2667-81.

Meng, F., Wiener, M.C., Sachs, J.R., Burns, C., Verma, P., Paweletz, C.P., Mazur, M.T., Deyanova, E.G., Yates, N.A. and Hendrickson, R.C. (**2007**). "Quantitative analysis of complex peptide mixtures using FTMS and differential mass spectrometry." *J Am Soc Mass Spectrom* 18(2): 226-33.

Mirgorodskaya, E., Braeuer, C., Fucini, P., Lehrach, H. and Gobom, J. (**2005**). "Nanoflow liquid chromatography coupled to matrix-assisted laser desorption/ionization mass spectrometry: sample preparation, data analysis, and application to the analysis of complex peptide mixtures." *Proteomics* 5(2): 399-408.

Mirzaei, H., Mcbee, J., Watts, J. and Aebersold, R. (**2007**). "Comparative evaluation of current peptide production platforms used in absolute quantification in proteomics." *Mol Cell Proteomics*.

Monigatti, F. and Berndt, P. (**2005**). "Algorithm for accurate similarity measurements of peptide mass fingerprints and its application." *J Am Soc Mass Spectrom* 16: 13-21.

Nesvizhskii, A.I., Vitek, O. and Aebersold, R. (**2007**). "Analysis and validation of proteomic data generated by tandem mass spectrometry." *Nat Methods* 4(10): 787-97.

Nie, L., Wu, G. and Zhang, W. (**2006**). "Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis." *Genetics* 174(4): 2229-43.

Nissum, M., Kuhfuss, S., Hauptmann, M., Obermaier, C., Sukop, U., Wildgruber, R., Weber, G., Eckerskorn, C. and Malmstrom, J. (**2007**). "Two-dimensional separation of human plasma proteins using iterative free-flow electrophoresis." *Proteomics* 7(23): 4218-27.

Noga, M.J., Asperger, A. and Silberring, J. (**2006**). "N-terminal $H_3/D_3$-acetylation for improved high-throughput peptide sequencing by matrix-assisted laser desorption/ionization mass spectrometry with a time-of-flight/time-of-flight analyzer." *Rapid Commun Mass Spectrom* 20(12): 1823-7.

Old, W.M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K.G., Mendoza, A., Sevinsky, J.R., Resing, K.A. and Ahn, N.G. (**2005**). "Comparison of label-free methods for quantifying human proteins by shotgun proteomics." *Mol Cell Proteomics* 4(10): 1487-502.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A. and Mann, M. (**2002**). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." *Mol Cell Proteomics* 1(5): 376-86.

Ong, S.E. and Mann, M. (**2005**). "Mass spectrometry-based proteomics turns quantitative." *Nat Chem Biol* 1(5): 252-62.

Pan, S., Zhang, H., Rush, J., Eng, J., Zhang, N., Patterson, D., Comb, M.J. and Aebersold, R. (**2005**). "High throughput proteome screening for biomarker detection." *Mol Cell Proteomics* 4(2): 182-90.

Pandhal, J., Wright, P.C. and Biggs, C.A. (**2007**). "A quantitative proteomic analysis of light adaptation in a globally significant marine cyanobacterium Prochlorococcus marinus MED4." *J Proteome Res* 6(3): 996-1005.

Papageorgopoulos, C., Caldwell, K., Shackleton, C., Schweingrubber, H. and Hellerstein, M.K. (**1999**). "Measuring protein synthesis by mass isotopomer distribution analysis (MIDA)." *Anal Biochem* 267(1): 1-16.

Pappin, D.J., Hojrup, P. and Bleasby, A.J. (**1993**). "Rapid identification of proteins by peptide-mass fingerprinting." *Curr Biol* 3(6): 327-32.

Paul, W. (**1990**). "Electromagnetic traps for charged and neutral particles (Nobel lecture)." *Angewandte Chemie International Edition in english* 29: 739-748.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (**1999**). "Probability-based protein identification by searching sequence databases using mass spectrometry data." *Electrophoresis* 20(18): 3551-67.

Plavina, T., Wakshull, E., Hancock, W.S. and Hincapie, M. (**2007**). "Combination of abundant protein depletion and multi-lectin affinity chromatography (M-LAC) for plasma protein biomarker discovery." *J Proteome Res* 6(2): 662-71.

Prakash, A., Mallick, P., Whiteaker, J., Zhang, H., Paulovich, A., Flory, M., Lee, H., Aebersold, R. and Schwikowski, B. (**2006**). "Signal maps for mass spectrometry-based comparative proteomics." *Mol Cell Proteomics* 5(3): 423-32.

Prakash, A., Piening, B., Whiteaker, J., Zhang, H., Shaffer, S.A., Martin, D., Hohmann, L., Cooke, K., Olson, J.M., Hansen, S., Flory, M.R., Lee, H., Watts, J., Goodlett, D.R., Aebersold, R., Paulovich, A. and Schwikowski, B. (**2007**). "Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics." *Mol Cell Proteomics* 6(10): 1741-8.

Pratt, J.M., Petty, J., Riba-Garcia, I., Robertson, D.H., Gaskell, S.J., Oliver, S.G. and Beynon, R.J. (**2002**). "Dynamics of protein turnover, a missing dimension in proteomics." *Mol Cell Proteomics* 1(8): 579-91.

Pratt, J.M., Robertson, D.H., Gaskell, S.J., Riba-Garcia, I., Hubbard, S.J., Sidhu, K., Oliver, S.G., Butler, P., Hayes, A., Petty, J. and Beynon, R.J. (**2002**). "Stable isotope labelling *in vivo* as an aid to protein identification in peptide mass fingerprinting." *Proteomics* 2(2): 157-63.

Pratt, J.M., Simpson, D.M., Doherty, M.K., Rivers, J., Gaskell, S.J. and Beynon, R.J. (**2006**). "Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes." *Nat Protoc* 1(2): 1029-43.

Righetti, P.G., Boschetti, E., Lomas, L. and Citterio, A. (**2006**). "Protein equalizer technology: The quest for a "democratic proteome"." *Proteomics* 6(14): 3980-92.

Righetti, P.G. and Boschetti, E. (**2007**). "Sherlock Holmes and the proteome-a detective story." *Febs J* 274(4): 897-905.

Rivers, J., Simpson, D.M., Robertson, D.H., Gaskell, S.J. and Beynon, R.J. (**2007**). "Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT." *Mol Cell Proteomics* 6(8): 1416-27.

Rivers, J., McDonald, L., Edwards, I.J., Beynon, R.J. (**2008**). "Asaparagine deamidation and the role of higher order protein structure." *Journal of Proteome Research*: in press.

Robinson, A.B. and Rudd, C.J. (**1974**). "Deamidation of glutaminyl and asparaginyl residues in peptides and proteins." *Curr Top Cell Regul* 8(0): 247-95.

Robinson, N.E., Robinson, A.B. and Merrifield, R.B. (**2001**). "Mass spectrometric evaluation of synthetic peptides as primary structure models for peptide and protein deamidation." *J Pept Res* 57(6): 483-93.

Robinson, N.E., Robinson, Z.W., Robinson, B.R., Robinson, A.L., Robinson, J.A., Robinson, M.L. and Robinson, A.B. (**2004**). "Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides." *J Pept Res* 63(5): 426-36.

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. and Pappin, D.J. (**2004**). "Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents." *Mol Cell Proteomics* 3(12): 1154-69.

Russell, W.K., Park, Z.Y. and Russell, D.H. (**2001**). "Proteolysis in mixed organic-aqueous solvent systems: applications for peptide mass mapping using mass spectrometry." *Anal Chem* 73(11): 2682-5.

Samuelsson, J., Dalevi, D., Levander, F. and Rognvaldsson, T. (**2004**). "Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting." *Bioinformatics* 20(18): 3628-35.

Schechter, I. and Berger, A. (**1967**). "On the size of the active site in proteases. I. Papain." *Biochem Biophys Res Commun* 27(2): 157-62.

Schmidt, A., Kellermann, J. and Lottspeich, F. (**2005**). "A novel strategy for quantitative proteomics using isotope-coded protein labels." *Proteomics* 5(1): 4-15.

Schnolzer, M., Jedrzejewski, P. and Lehmann, W.D. (**1996**). "Protease-catalyzed incorporation of $^{18}O$ into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry." *Electrophoresis* 17(5): 945-53.

Scigelova, M., Makarov, A. (**2006**). "Orbitrap mass analyzer - overview and applications in proteomics." *practical proteomics* 1: 16-21.

Sebastiano, R., Citterio, A., Lapadula, M. and Righetti, P.G. (**2003**). "A new deuterated alkylating agent for quantitative proteomics." *Rapid Communications in Mass Spectrometry* 17(21): 2380-6.

Sechi, S. and Chait, B.T. (**1998**). "Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification." *Anal Chem* 70(24): 5150-8.

Seedorf, H., Fricke, W.F., Veith, B., Bruggemann, H., Liesegang, H., Strittmatter, A., Miethke, M., Buckel, W., Hinderberger, J., Li, F., Hagemeier, C., Thauer, R.K. and Gottschalk, G. (**2008**). "The genome of Clostridium kluyveri, a strict anaerobe with unique metabolic features." *Proc Natl Acad Sci U S A.*

Sennels, L., Salek, M., Lomas, L., Boschetti, E., Righetti, P.G. and Rappsilber, J. (**2007**). "Proteomic analysis of human blood serum using peptide library beads." *J Proteome Res* 6(10): 4055-62.

Siepen, J.A., Keevil, E.J., Knight, D. and Hubbard, S.J. (**2007**). "Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics." *J Proteome Res* 6(1): 399-408.

Siuti, N. and Kelleher, N.L. (**2007**). "Decoding protein modifications using top-down mass spectrometry." *Nat Methods* 4(10): 817-21.

Smith, R.D., Anderson, G.A., Lipton, M.S., Pasa-Tolic, L., Shen, Y., Conrads, T.P., Veenstra, T.D. and Udseth, H.R. (**2002**). "An accurate mass tag strategy for quantitative and high-throughput proteome measurements." *Proteomics* 2(5): 513-23.

Snijders, A.P., De Vos, M.G., De Koning, B. and Wright, P.C. (**2005**). "A fast method for quantitative proteomics based on a combination between two-dimensional electrophoresis and $^{15}N$-metabolic labelling." *Electrophoresis* 26(16): 3191-9.

Snijders, A.P., De Vos, M.G. and Wright, P.C. (**2005**). "Novel approach for peptide quantitation and sequencing based on $^{15}N$ and $^{13}C$ metabolic labeling." *J Proteome Res* 4(2): 578-85.

Snijders, A.P., De Koning, B. and Wright, P.C. (**2007**). "Relative quantification of proteins across the species boundary through the use of shared peptides." *J Proteome Res* 6(1): 97-104.

Spengler, B., Kirsch, D., Kaufmann, R. and Jaeger, E. (**1992**). "Peptide sequencing by matrix-assisted laser-desorption mass spectrometry." *Rapid Commun Mass Spectrom* 6(2): 105-8.

Stapels, M.D. and Barofsky, D.F. (**2004**). "Complementary use of MALDI and ESI for the HPLC-MS/MS analysis of DNA-binding proteins." *Anal Chem* 76(18): 5423-30.

Stapels, M.D., Cho, J.C., Giovannoni, S.J. and Barofsky, D.F. (**2004**). "Proteomic analysis of novel marine bacteria using MALDI and ESI mass spectrometry." *J Biomol Tech* 15(3): 191-8.

Steen, H. and Mann, M. (**2004**). "The ABC's (and XYZ's) of peptide sequencing." *Nat Rev Mol Cell Biol* 5(9): 699-711.

Stephens, W.E. (**1946**). "A pulsed mass spectrometer with time dispersion." *Physical Review* 69: 691.

Strader, M.B., Tabb, D.L., Hervey, W.J., Pan, C. and Hurst, G.B. (**2006**). "Efficient and specific trypsin digestion of microgram to nanogram quantities of proteins in organic-aqueous solvent systems." *Anal Chem* 78(1): 125-34.

Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J. and Hunt, D.F. (**2004**). "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry." *Proc Natl Acad Sci U S A* 101(26): 9528-33.

Tabb, D.L., Mcdonald, W.H. and Yates, J.R., 3rd (**2002**). "DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics." *J Proteome Res* 1(1): 21-6.

Thulasiraman, V., Lin, S., Gheorghiu, L., Lathrop, J., Lomas, L., Hammond, D. and Boschetti, E. (**2005**). "Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands." *Electrophoresis* 26(18): 3561-71.

Turck, C.W., Falick, A.M., Kowalak, J.A., Lane, W.S., Lilley, K.S., Phinney, B.S., Weintraub, S.T., Witkowska, H.E. and Yates, N.A. (**2007**). "The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation." *Mol Cell Proteomics* 6(8): 1291-8.

Turner, B.M. (**2001**). Chapter 4: Histone tails: Modifications and Epigenetic Information. Oxford, Blackwell Science Ltd.

Unlu, M., Morgan, M.E. and Minden, J.S. (**1997**). "Difference gel electrophoresis: a single gel method for detecting changes in protein extracts." *Electrophoresis* 18(11): 2071-7.

Vincent, S.G., Cunningham, P.R., Stephens, N.L., Halayko, A.J. and Fisher, J.T. (**1997**). "Quantitative densitometry of proteins stained with coomassie blue using a Hewlett Packard scanjet scanner and Scanplot software." *Electrophoresis* 18(1): 67-71.

Viswanathan, S., Unlu, M. and Minden, J.S. (**2006**). "Two-dimensional difference gel electrophoresis." *Nat Protoc* 1(3): 1351-8.

Wang, M.Z. and Fitzgerald, M.C. (**2001**). "A solid sample preparation method that reduces signal suppression effects in the MALDI analysis of peptides." *Anal Chem* 73(3): 625-31.

Wang, G., Wu, W.W., Zeng, W., Chou, C.L. and Shen, R.F. (**2006**). "Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes." *J Proteome Res* 5(5): 1214-23.

Washburn, M.P., Wolters, D. and Yates, J.R., 3rd (**2001**). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." *Nat Biotechnol* 19(3): 242-7.

Weinkam, R.J., Wen, J.H., Furst, D.E. and Levin, V.A. (**1978**). "Analysis for 1,3-bis(2-chloroethyl)-1-nitrosourea by chemical ionization mass spectrometry." *Clin Chem* 24(1): 45-9.

Weintraub, S.J. and Manson, S.R. (**2004**). "Asparagine deamidation: a regulatory hourglass." *Mech Ageing Dev* 125(4): 255-7.

Wiener, M.C., Sachs, J.R., Deyanova, E.G. and Yates, N.A. (**2004**). "Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures." *Anal Chem* 76(20): 6085-96.

Wienkoop, S. and Weckwerth, W. (**2006**). "Relative and absolute quantitative shotgun proteomics: targeting low-abundance proteins in Arabidopsis thaliana." *J Exp Bot* 57(7): 1529-35.

Wiese, S., Reidegeld, K.A., Meyer, H.E. and Warscheid, B. (**2007**). "Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research." *Proteomics* 7(3): 340-50.

Wong, J.W., Sullivan, M.J. and Cagney, G. (**2007**). "Computational methods for the comparative quantification of proteins in label-free $LC^n$-MS experiments." *Brief Bioinform.*

Wu, C., Robertson, D.H., Hubbard, S.J., Gaskell, S.J. and Beynon, R.J. (**1999**). "Proteolysis of native proteins. Trapping of a reaction intermediate." *J Biol Chem* 274(2): 1108-15.

Wu, C.C. and Yates, J.R., 3rd (**2003**). "The application of mass spectrometry to membrane proteomics." *Nat Biotechnol* 21(3): 262-7.

Wu, C.C., MacCoss, M.J., Howell, K.E., Matthews, D.E. and Yates, J.R., 3rd (**2004**). "Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis." *Anal Chem* 76(17): 4951-9.

Yang, J., Wongsa, S., Kadirkamanathan, V., Billings, S.A. and Wright, P.C. (**2005**). "Metabolic flux distribution analysis by $^{13}C$-tracer experiments using the Markov chain-Monte Carlo method." *Biochem Soc Trans* 33(Pt 6): 1421-2.

Yang, Y., Zhang, S., Howe, K., Wilson, D.B., Moser, F., Irwin, D. and Thannhauser, T.W. (**2007**). "A comparison of nLC-ESI-MS/MS and nLC-MALDI-MS/MS for GeLC-based protein identification and iTRAQ-based shotgun quantitative proteomics." *J Biomol Tech* 18(4): 226-37.

Yao, X., Freas, A., Ramirez, J., Demirev, P.A. and Fenselau, C. (**2001**). "Proteolytic $^{18}O$ labeling for comparative proteomics: model studies with two serotypes of adenovirus." *Anal Chem* 73(13): 2836-42.

Yates, J.R., 3rd, Eng, J.K., Mccormack, A.L. and Schieltz, D. (**1995**). "Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database." *Anal Chem* 67(8): 1426-36.

Yu, E.Z., Burba, A.E. and Gerstein, M. (**2007**). "PARE: a tool for comparing protein abundance and mRNA expression data." *BMC Bioinformatics* 8: 309.

Zhang, W. and Chait, B.T. (**2000**). "ProFound: an expert system for protein identification using mass spectrometric peptide mapping information." *Anal Chem* 72(11): 2482-9.

Zhang, R., Sioma, C.S., Wang, S. and Regnier, F.E. (**2001**). "Fractionation of isotopically labeled peptides in quantitative proteomics." *Anal Chem* 73(21): 5142-9.

Zhang, R. and Regnier, F.E. (**2002**). "Minimizing resolution of isotopically coded peptides in comparative proteomics." *J Proteome Res* 1(2): 139-47.

Zhang, B., Verberkmoes, N.C., Langston, M.A., Uberbacher, E., Hettich, R.L. and Samatova, N.F. (**2006**). "Detecting differential and correlated protein expression in label-free shotgun proteomics." *J Proteome Res* 5(11): 2909-18.

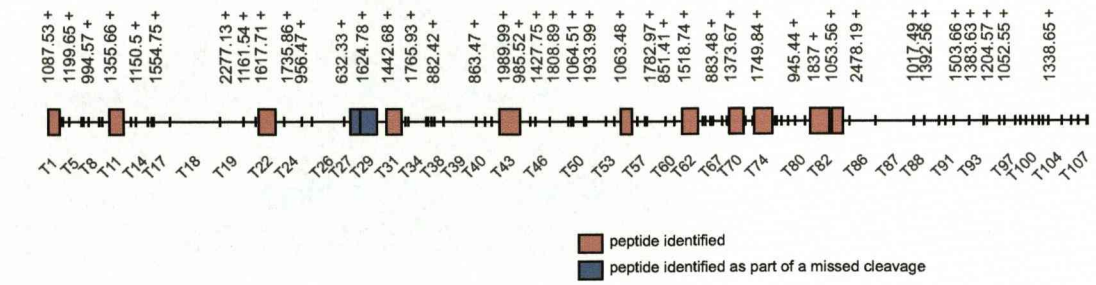
Zhang, H., Liu, A.Y., Loriaux, P., Wollscheid, B., Zhou, Y., Watts, J.D. and Aebersold, R. (**2007**). "Mass spectrometric detection of tissue proteins in plasma." *Mol Cell Proteomics* 6(1): 64-71.

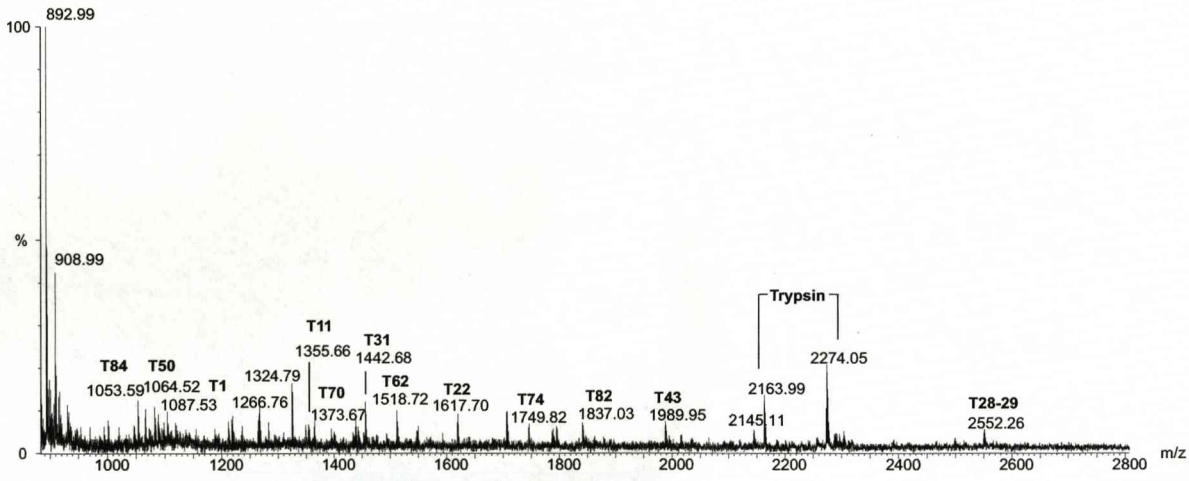
Zhu, H., Hunter, T.C., Pan, S., Yau, P.M., Bradbury, E.M. and Chen, X. (**2002**). "Residue-specific mass signatures for the efficient detection of protein modifications by mass spectrometry." *Anal Chem* 74(7): 1687-94.

Zolotarjova, N., Martosella, J., Nicol, G., Bailey, J., Boyes, B.E. and Barrett, W.C. (**2005**). "Differences among techniques for high-abundant protein depletion." *Proteomics* 5(13): 3304-13.

Zubarev, R.A.K., N. L. Mclafferty, F. W (**1998**). "Electron capture dissociation of multiply charged protein cations. A nonergodic process." *J Am Chem Soc* 120: 3265-3266.

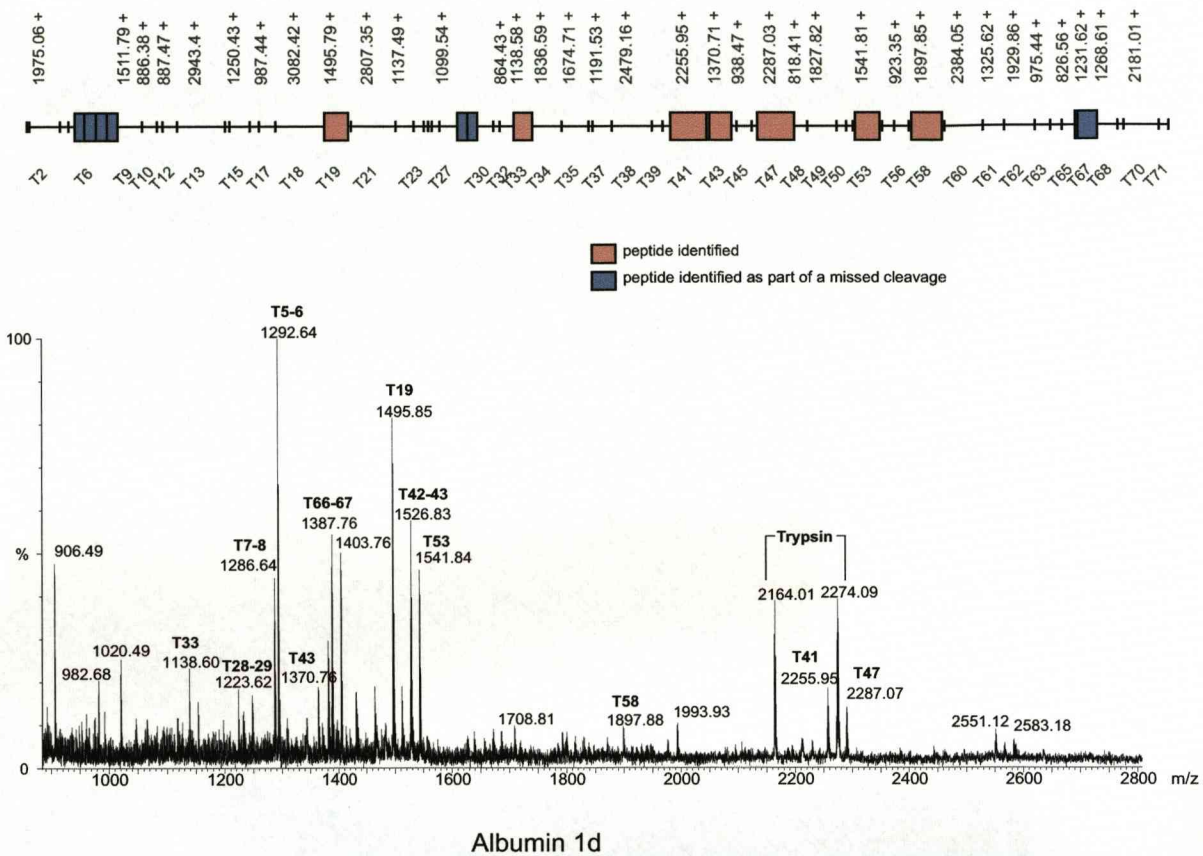
## 8. SUPPLEMENTARY FIGURES



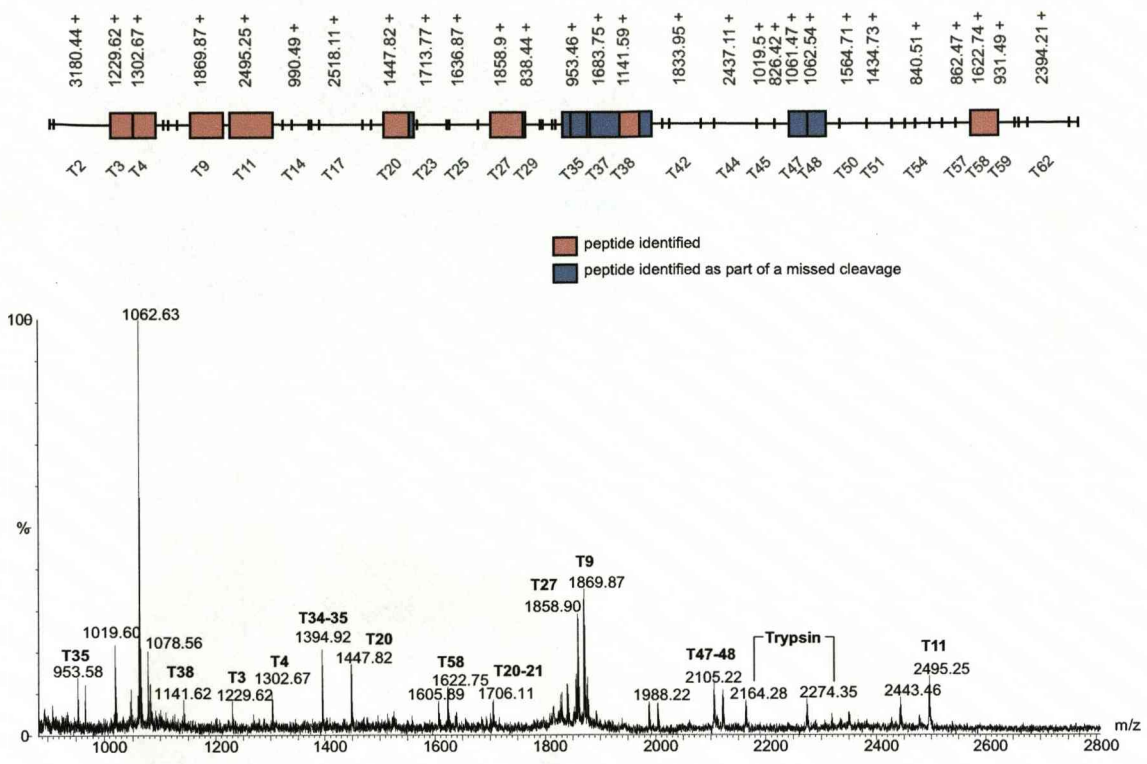

Glycogen phosphorylase 1d

Supplementary figure 1. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



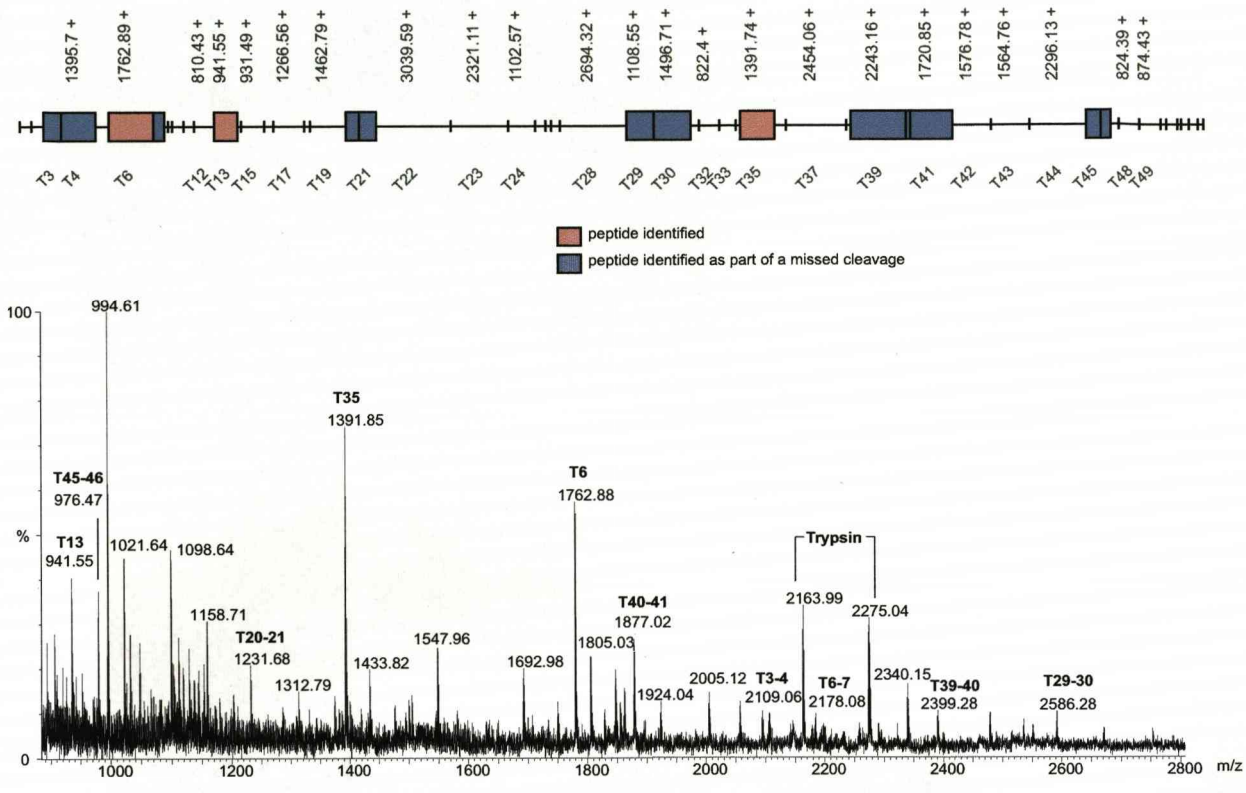

Albumin 1d

Supplementary figure 2. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



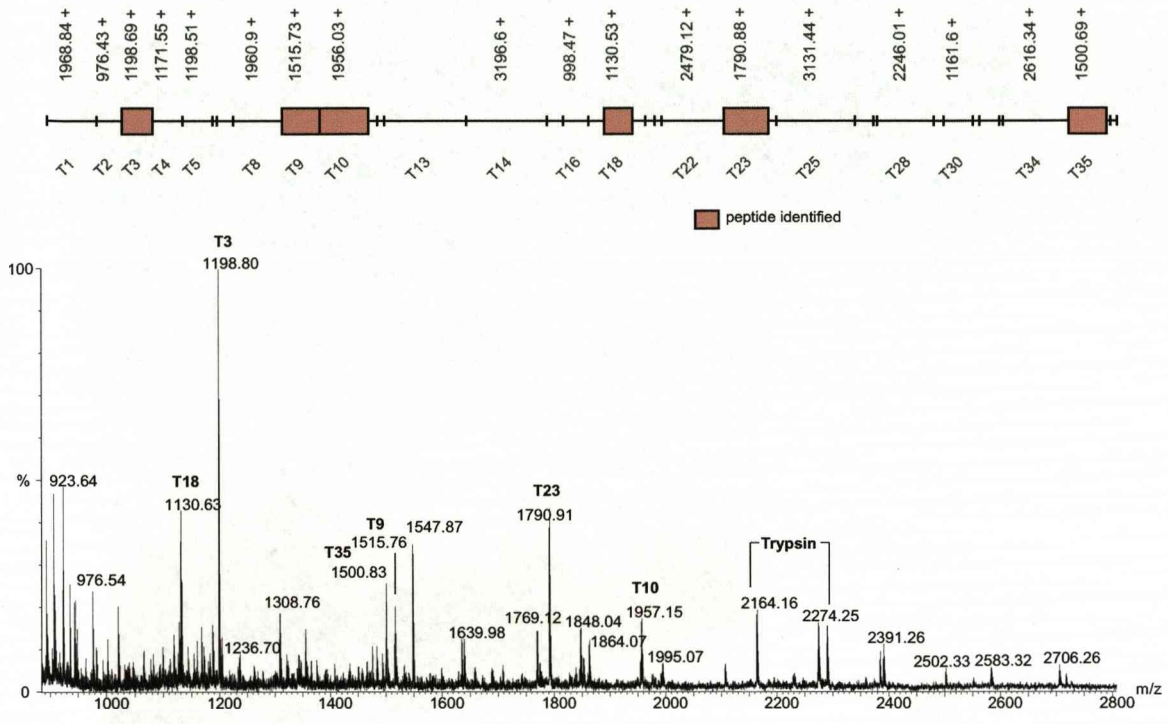

Pyruvate kinase 1d

Supplementary figure 3. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



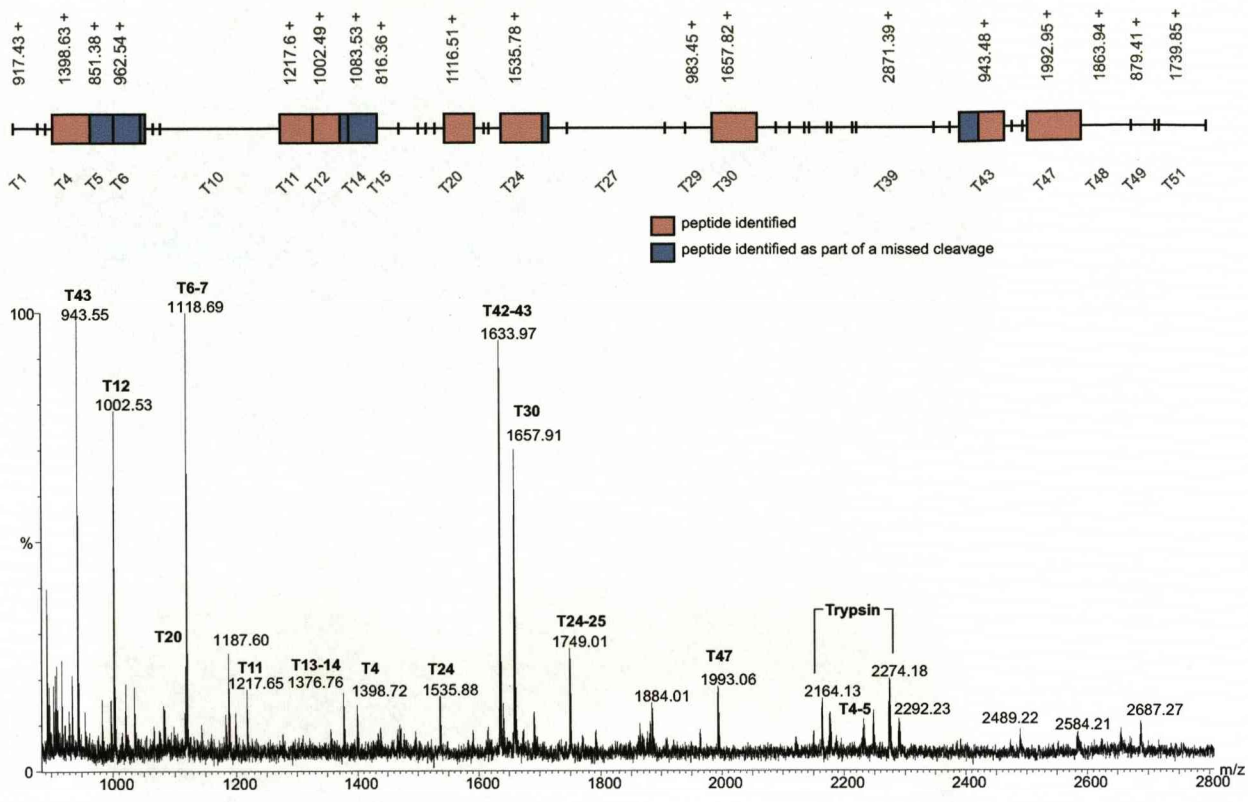

Beta enolase 1d

Supplementary figure 4. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



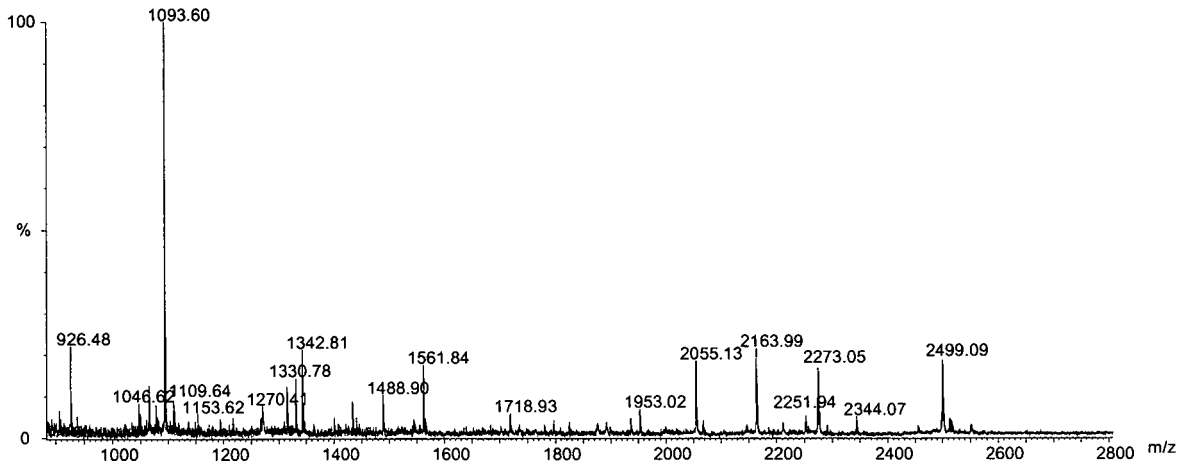

Alpha actin 1d

Supplementary figure 5. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



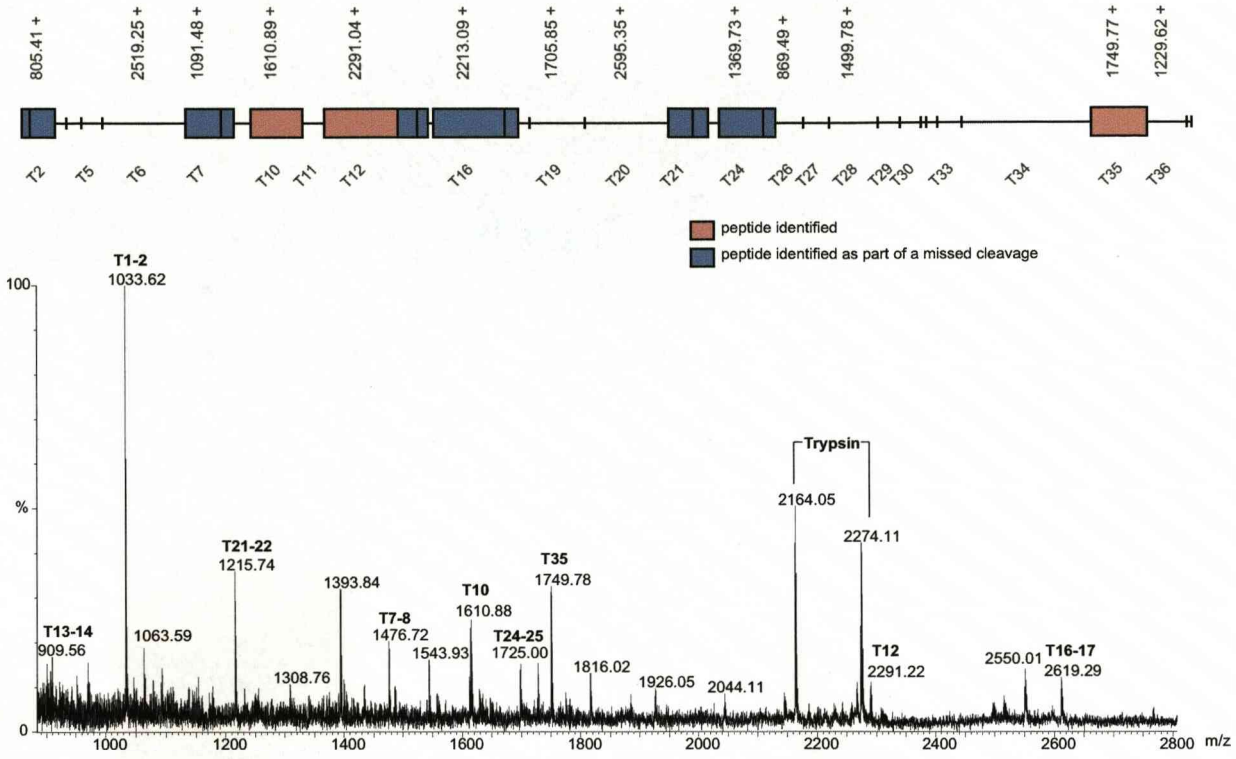

Creatine kinase 1d

Supplementary figure 6. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



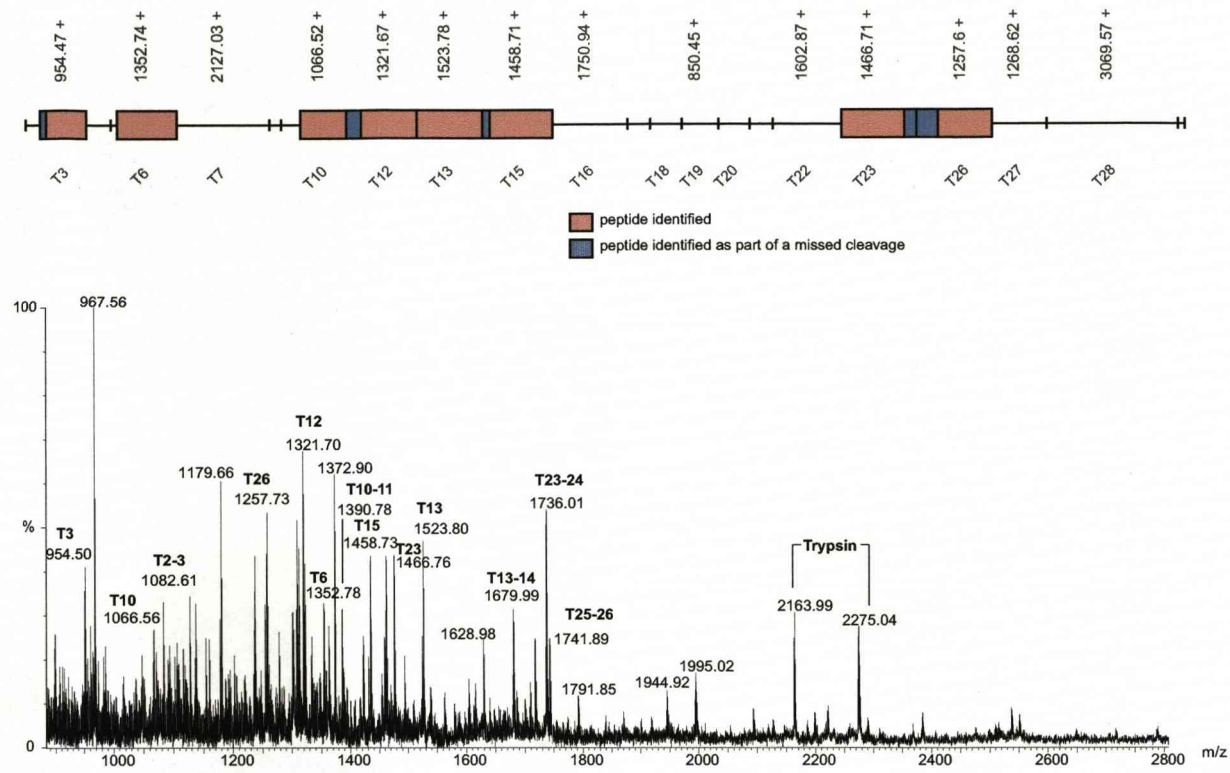

Aldolase A (muscle type) 1d

Supplementary figure 7. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



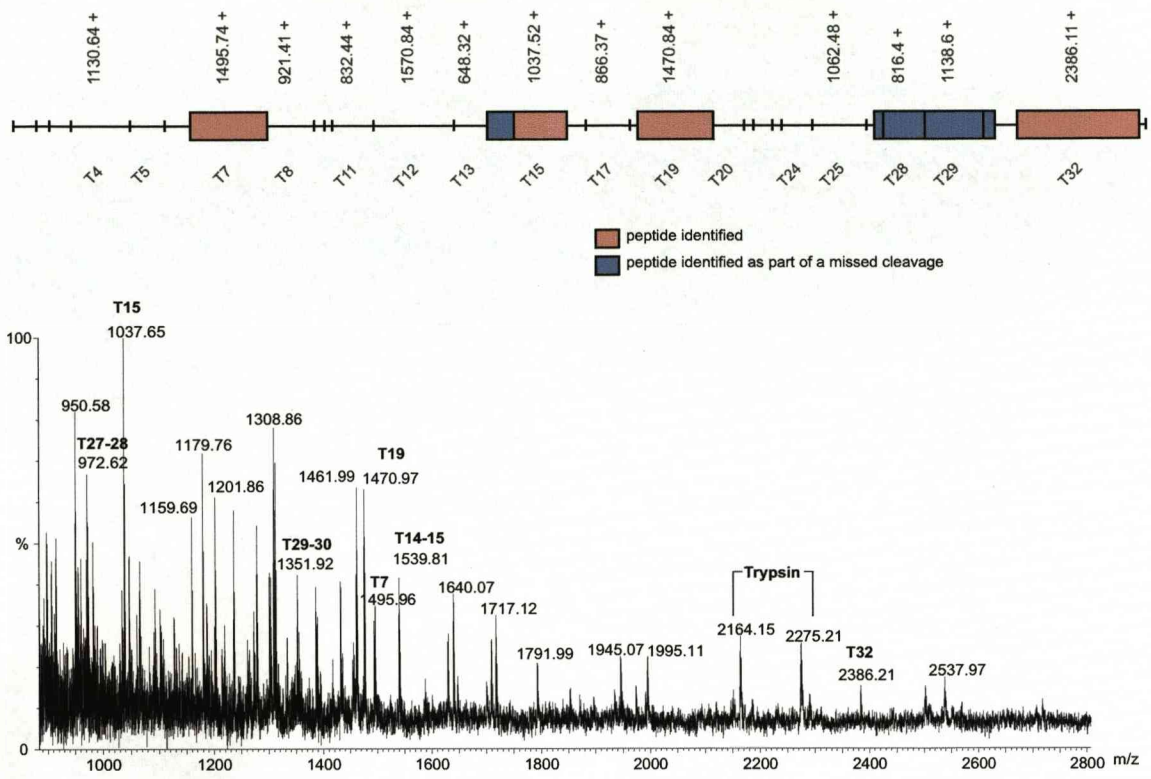

Glyceraldehyde 3-phosphate dehydrogenase 1d

Supplementary figure 8. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



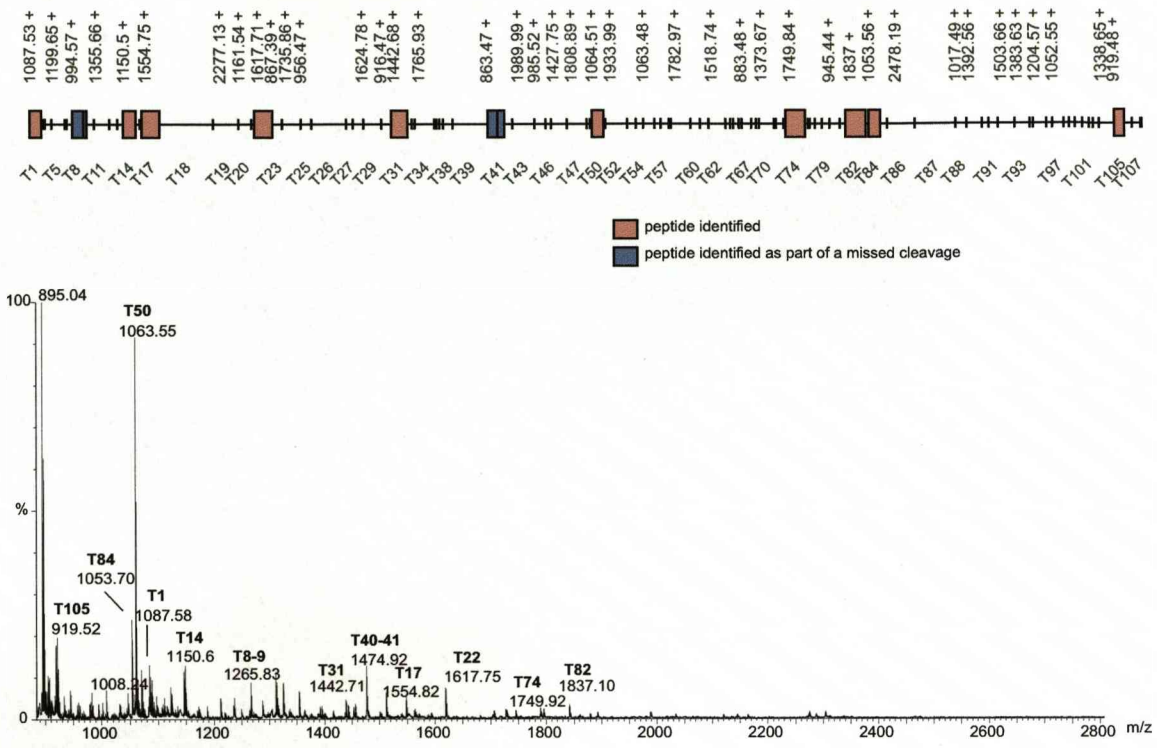

Triose phosphate isomerase 1d

Supplementary figure 9. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



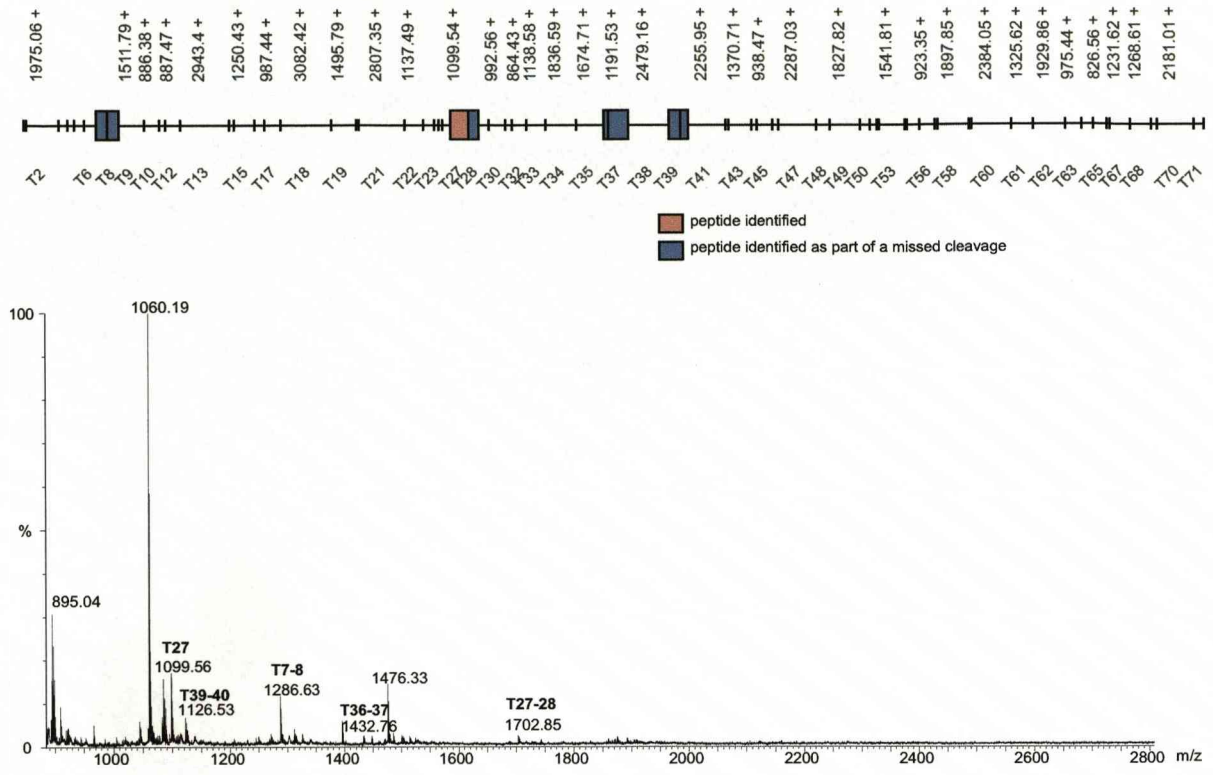

Adenylate kinase 1d

Supplementary figure 10. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



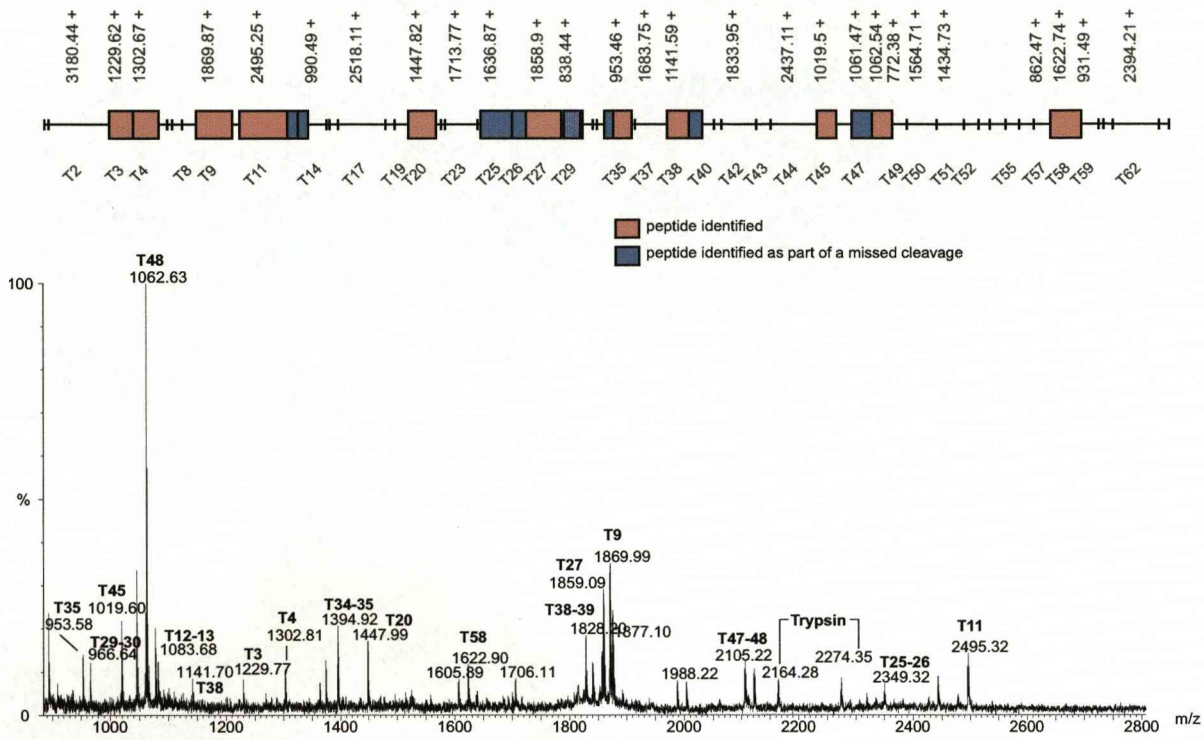

Glycogen phosphorylase 30d

Supplementary figure 11. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



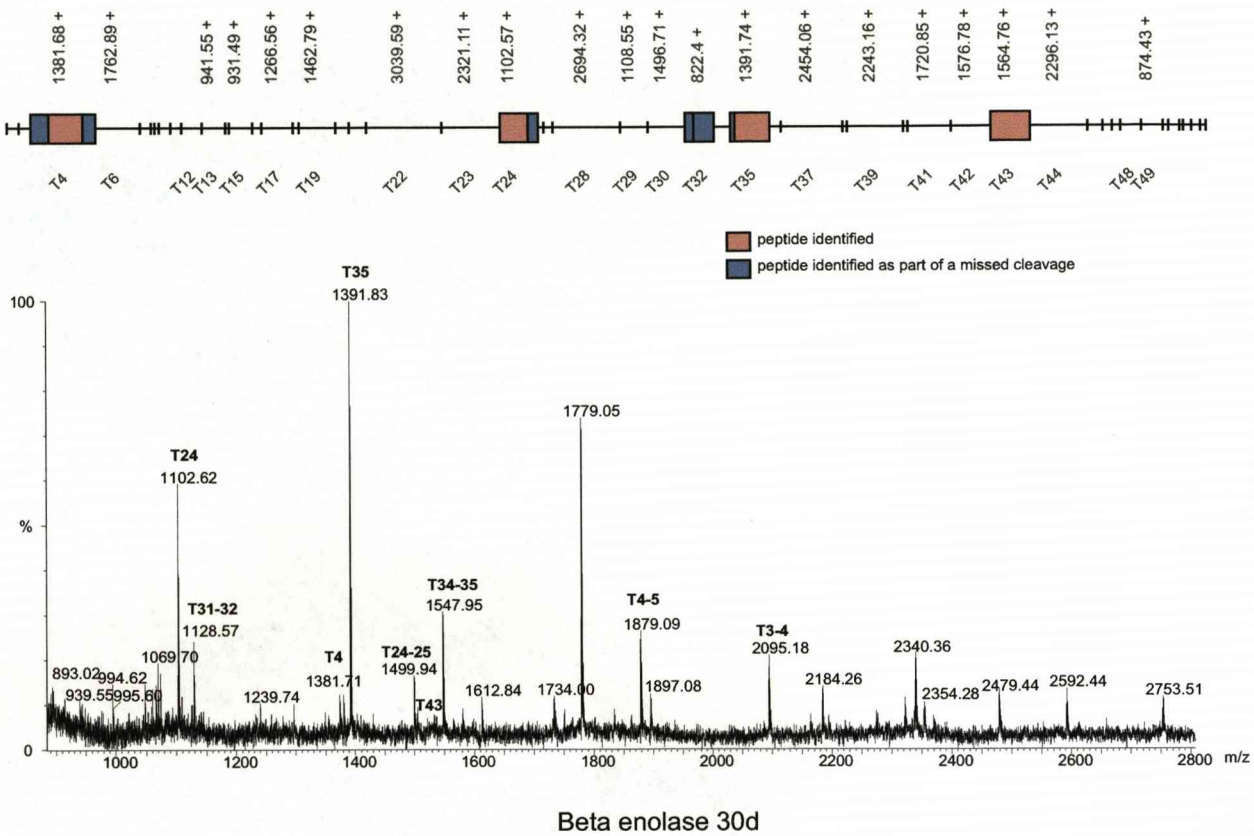

Albumin 30d

Supplementary figure 12. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



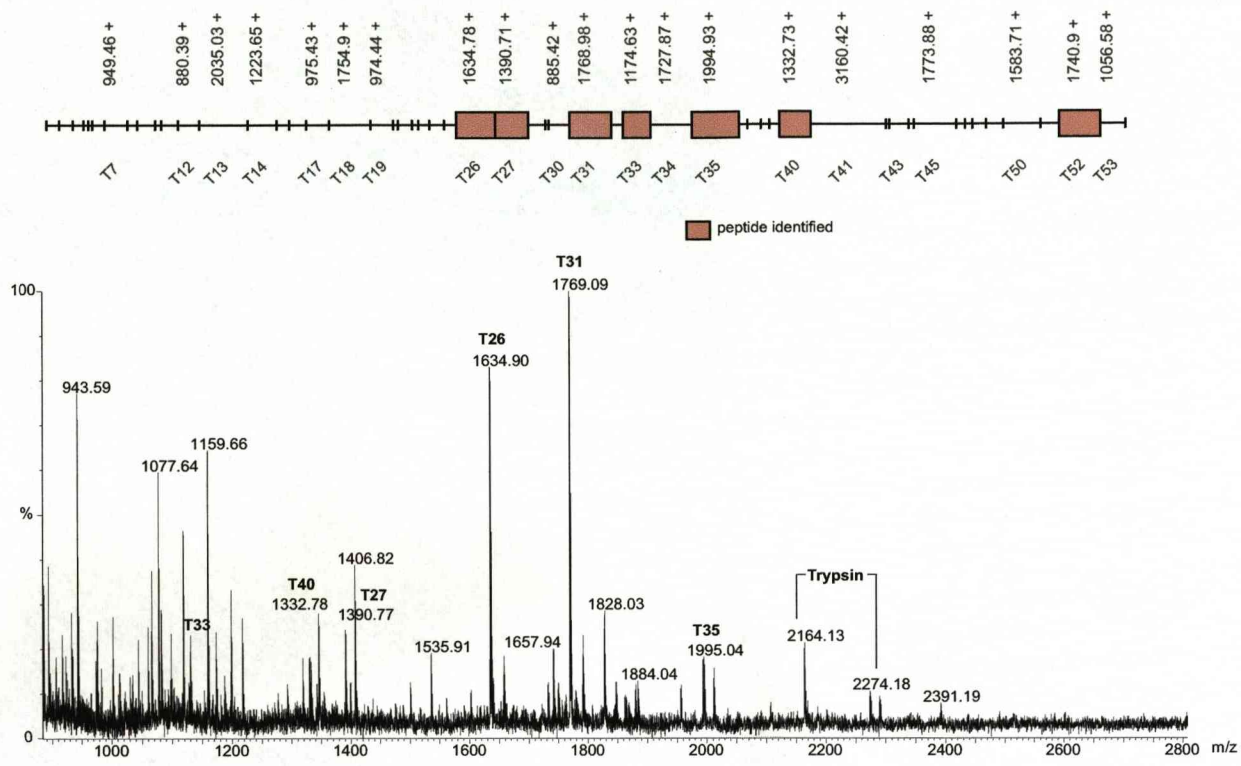

Pyruvate kinase 30d

Supplementary figure 13. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



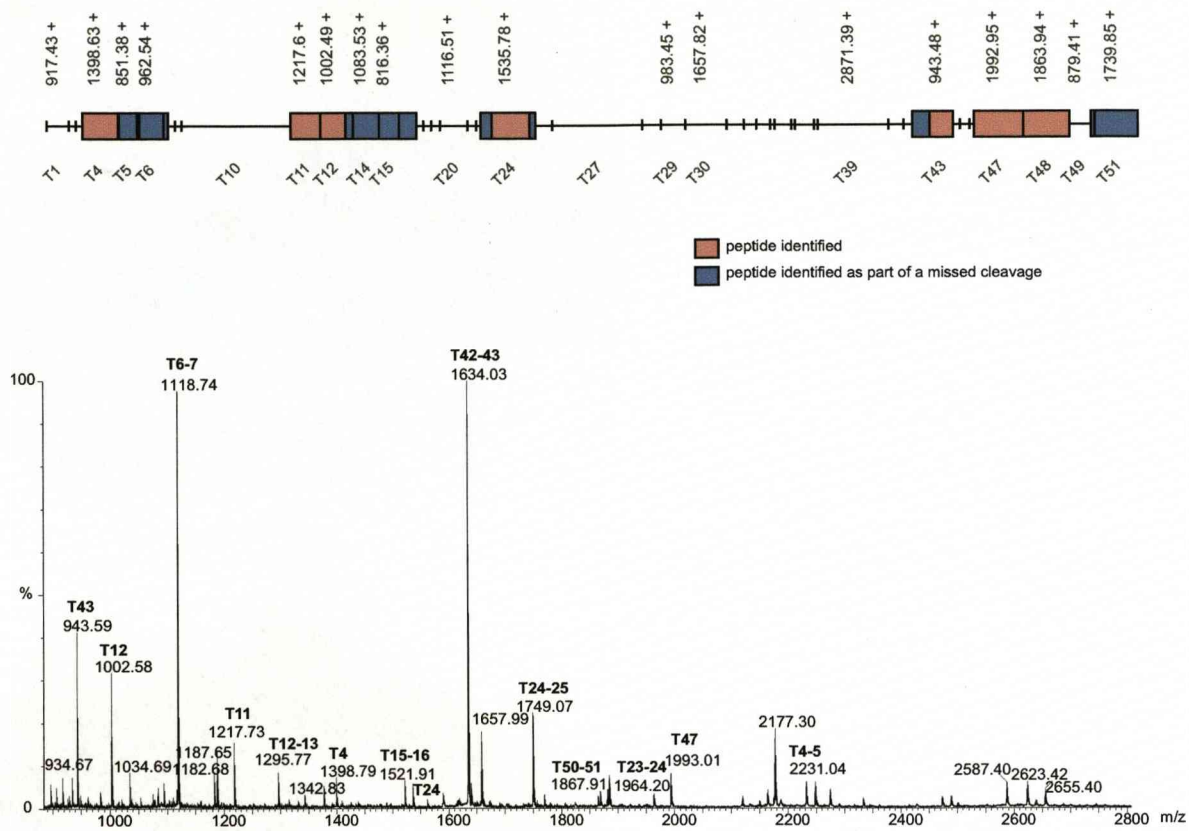

Beta enolase 30d

Supplementary figure 14. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



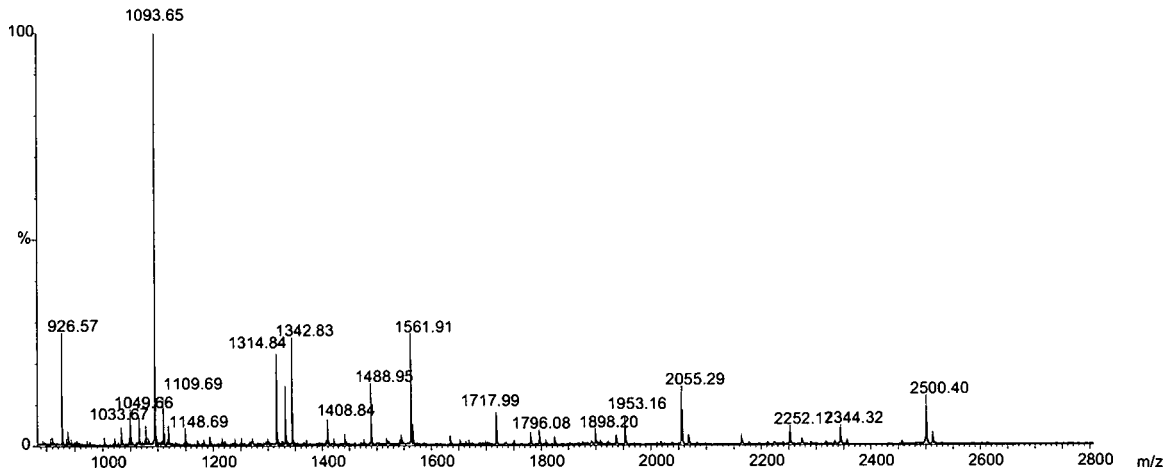

Phosphoglycerate kinase 30d

Supplementary figure 15. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



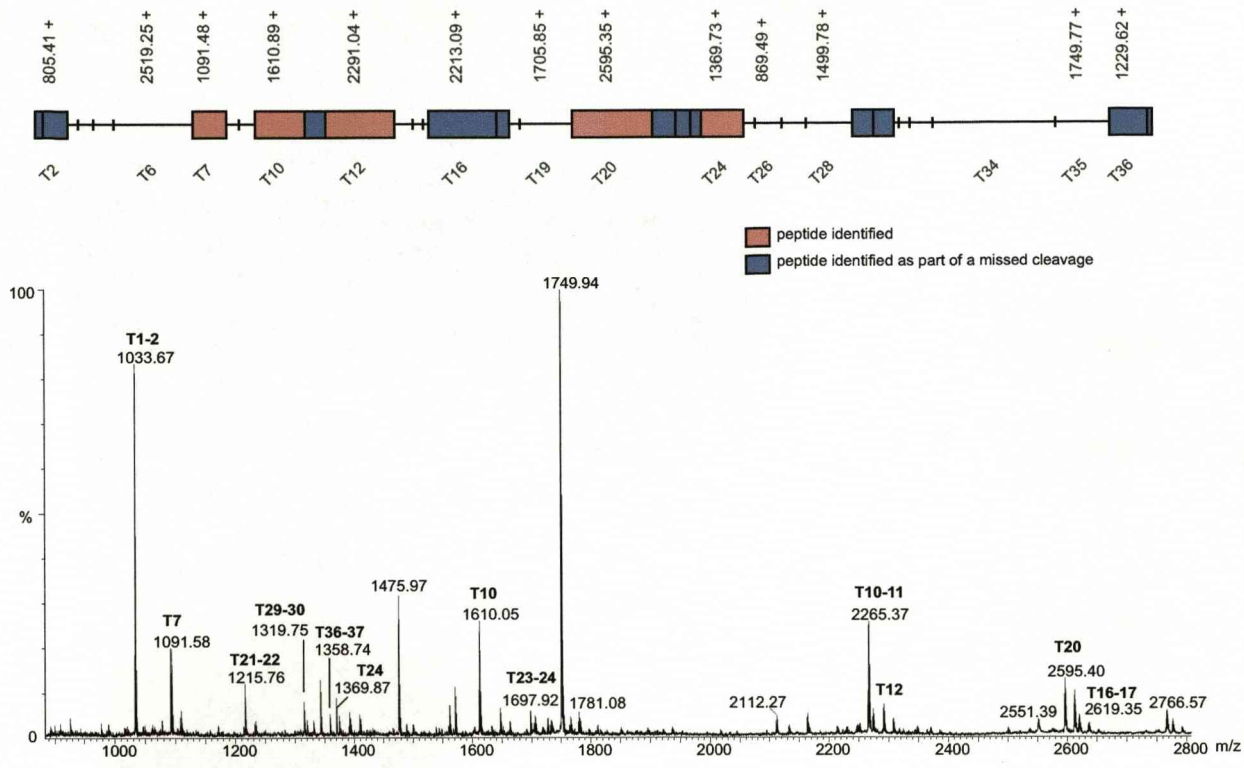

Creatine kinase 30d

Supplementary figure 16. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



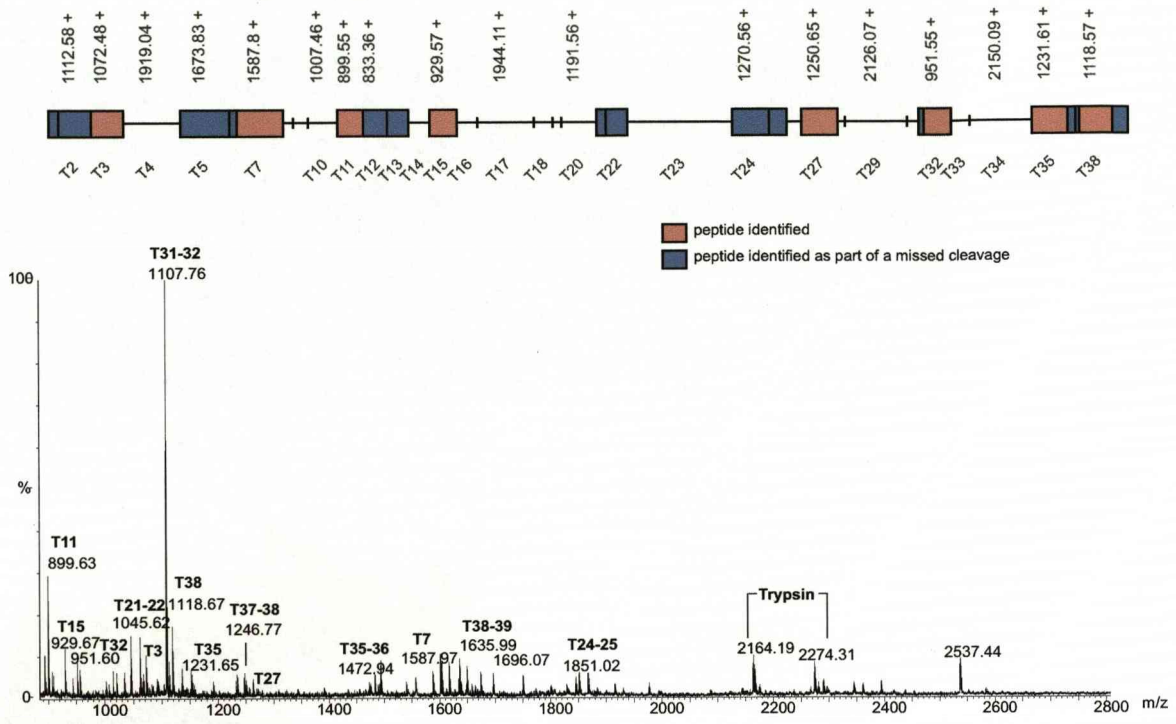

Aldolase A (muscle type) 30d

Supplementary figure 17. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



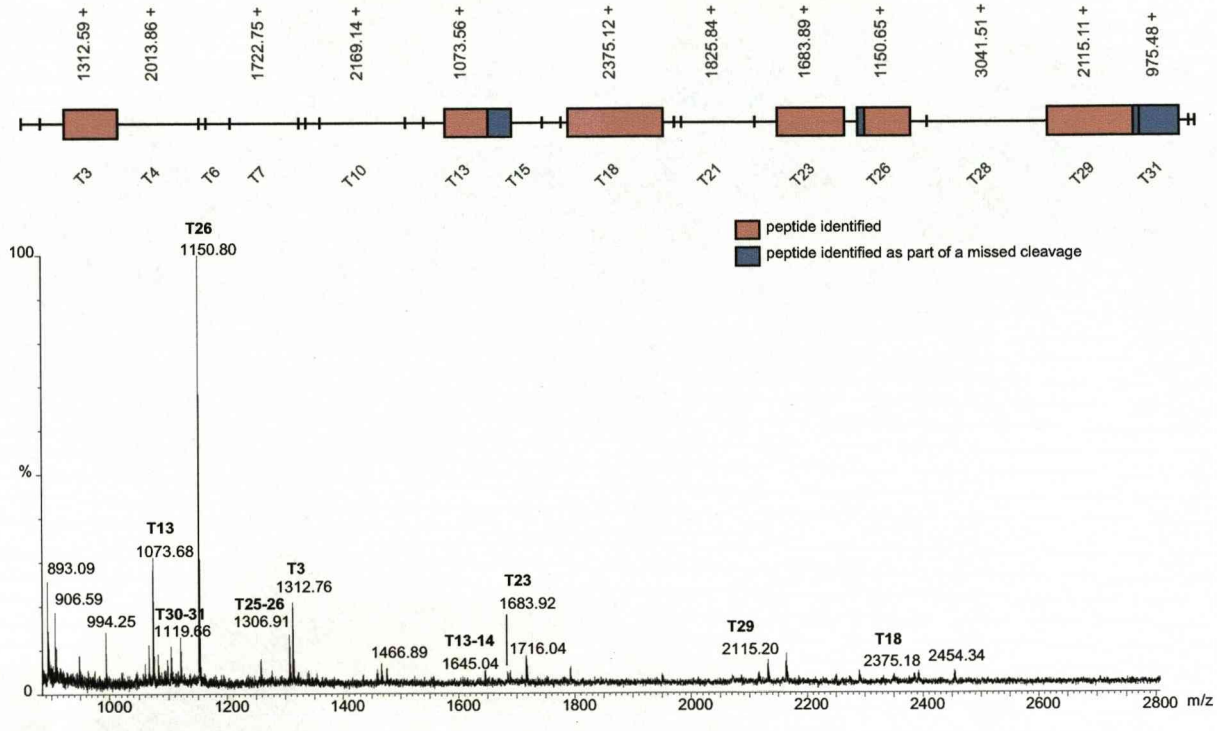

Glyceraldehyde 3-phosphate dehydrogenase 30d

Supplementary figure 18. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



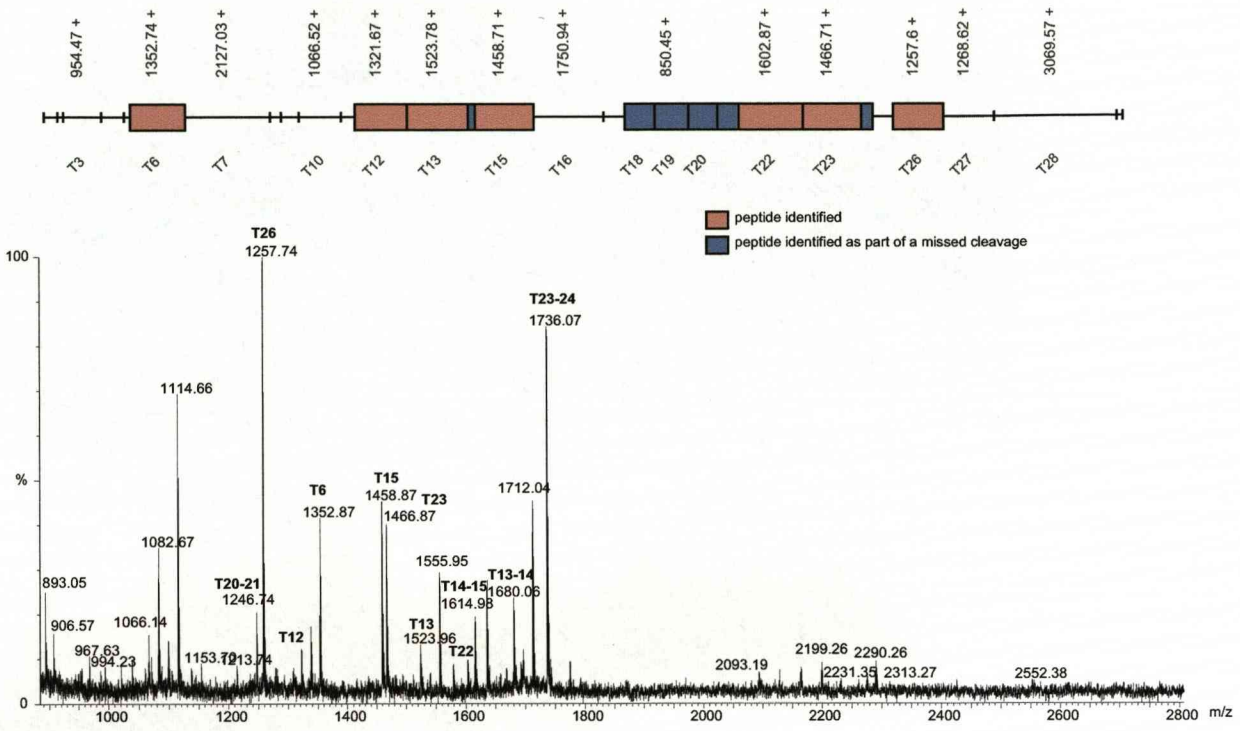

Lactate dehydrogenase 30d

Supplementary figure 19. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



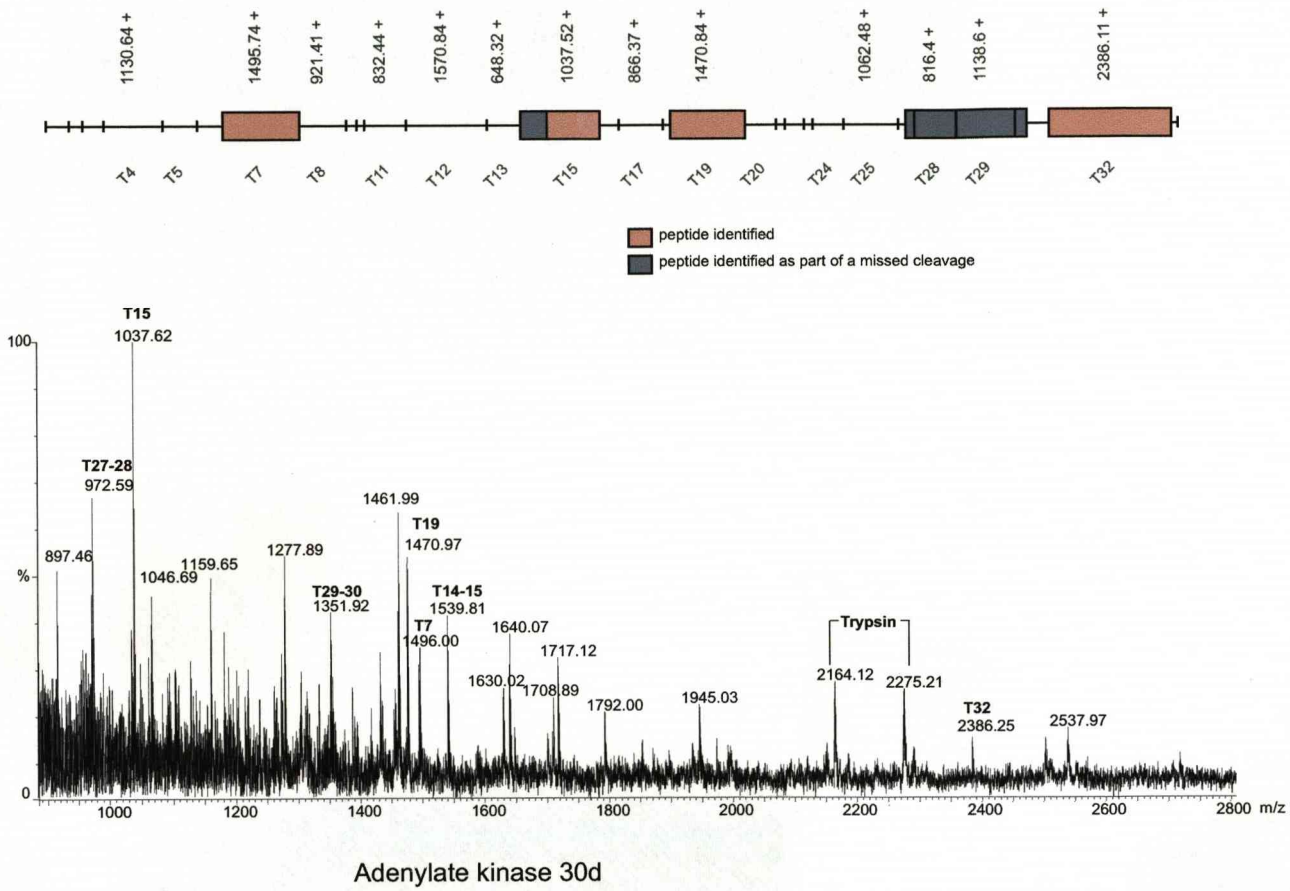

Phosphoglycerate mutase 30d

Supplementary figure 20. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.



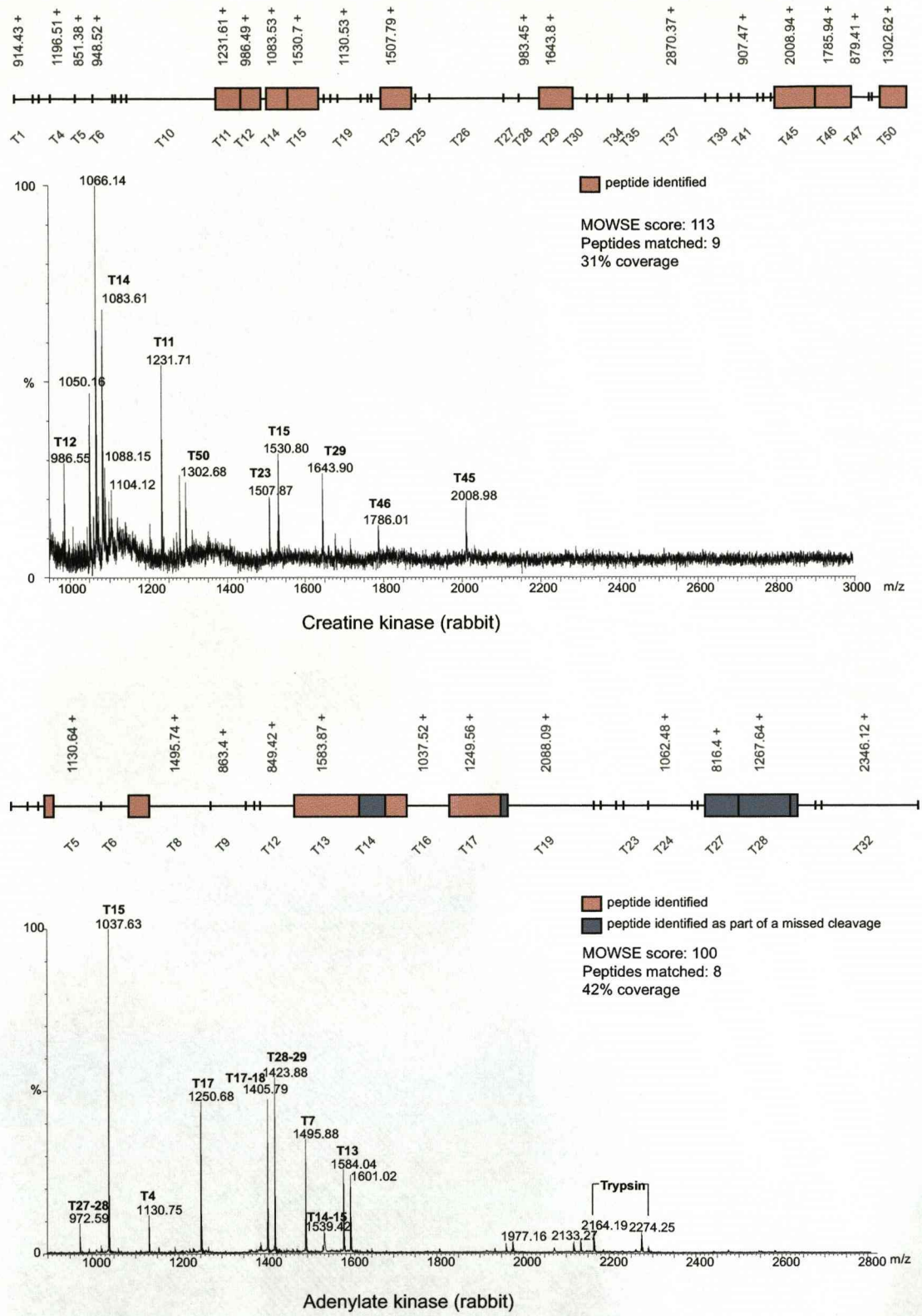

Triose phosphate isomerase 30d

Supplementary figure 21. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.

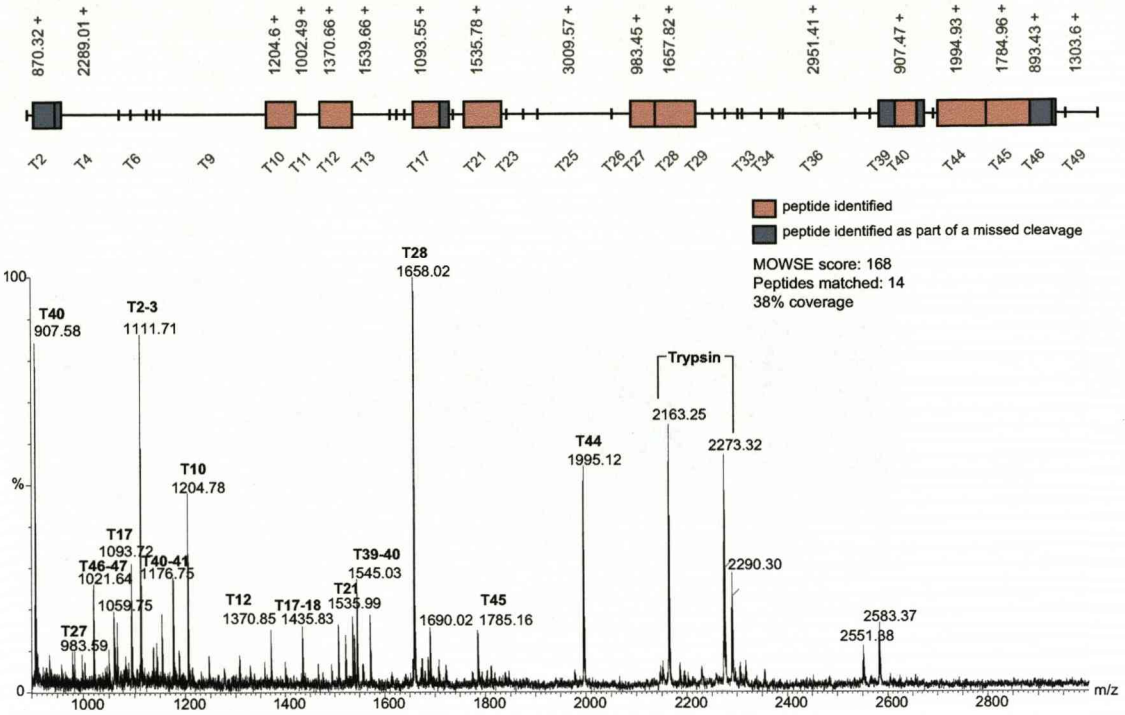



Adenylate kinase 30d

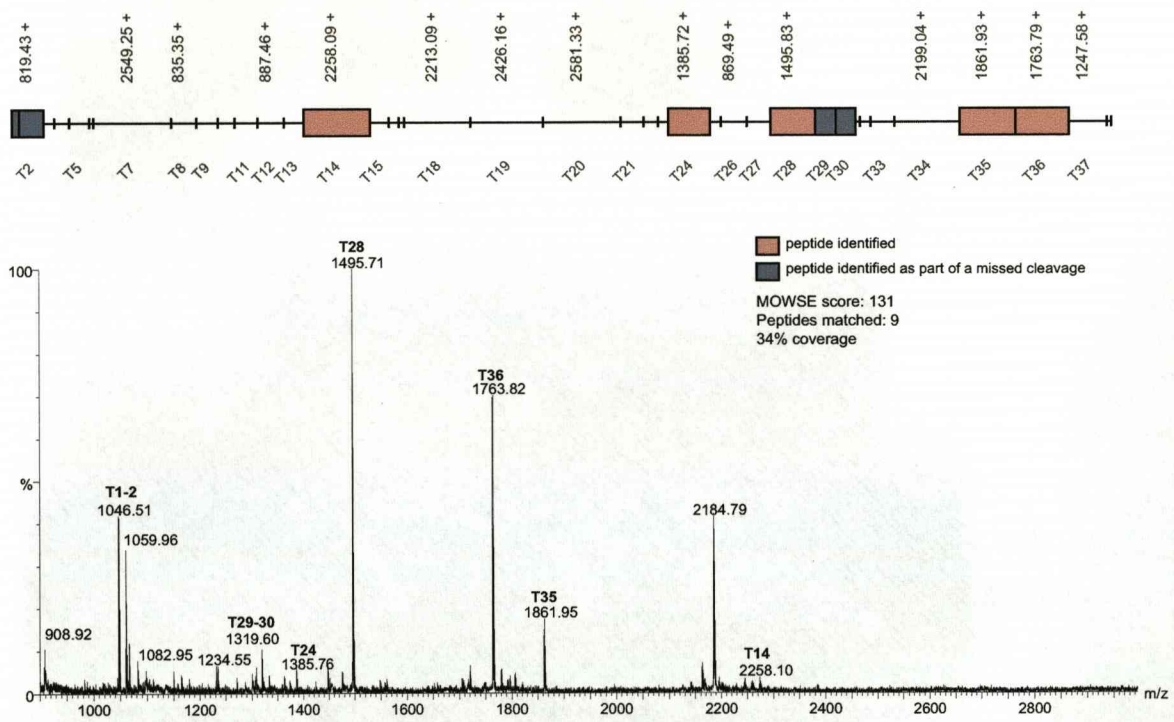
Supplementary figure 22. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 31.




Supplementary figure 23. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 73.

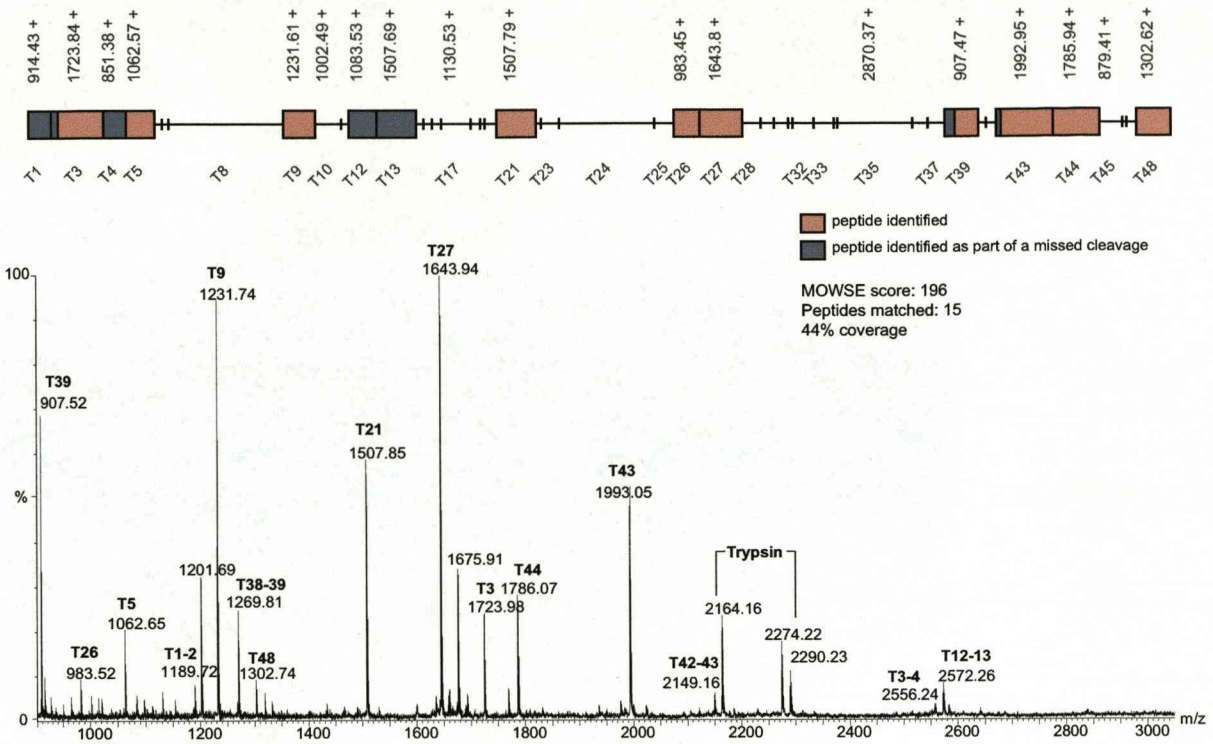



Creatine kinase (carp)

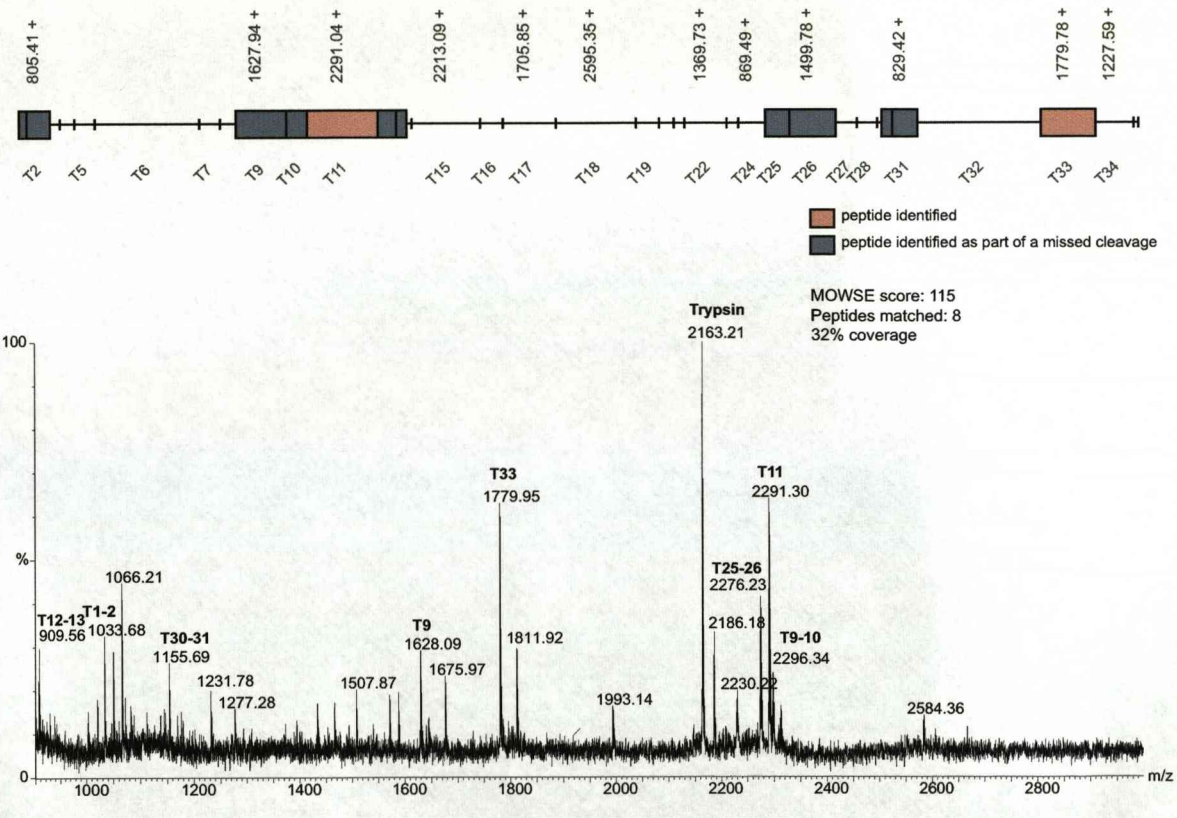



Glyceraldehyde 3-phosphate dehydrogenase (carp)

Supplementary figure 24. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 74.




Creatine kinase (mouse)




Glyceraldehyde 3-phosphate dehydrogenase (mouse)

Supplementary figure 25. Peptide map and annotated mass spectrum from in gel digestion and peptide mass fingerprinting by MALDI-ToF MS for protein identification, figure 74.

## 9. PUBLICATIONS AS A CONSEQUENCE OF THIS THESIS

2008 **Rivers J.** and Beynon R.J. (2008) 'Absolute protein quantification of normalisation' (in preparation for Proteomics)

Bogdan I., **Rivers J.**, Coca D. and Beynon R. J. (2008) 'High-performance hardware implementation of a parallel data-base search engine for real-time peptide mass fingerprinting' (submitted to Bioinformatics)

**Rivers J.**, McDonald L., Edwards I. J. and Beynon R. J. (2008) 'Asparagine deamidation and the role of higher order protein structure' Journal of Proteome Research 7 (3), 927-7

2007 **Rivers J.**, Simpson D. M., Robertson D. H. L., Gaskell S. J. and Beynon R. J. (2007) 'Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT' Molecular and Cellular Proteomics 6 (8), 1416 - 1427

Bogdan I., Coca D., **Rivers J.** and Beynon R. J. (2007) 'Hardware acceleration of processing of mass spectrometric data for Proteomics' Bioinformatics 23 (6), 724-731

2006 Pratt J. M., Simpson D. M., Doherty M. K., **Rivers J.**, Gaskell S. J. and Beynon R. J. (2006) 'Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes' Nature Protocols 3 (2), 1029 – 1043

Contribution to each publication:

**Rivers J.** and Beynon R.J. (2008) 'Absolute protein quantification of normalisation' (in preparation for Proteomics)

I was responsible for initial exploratory work, all subsequent experimental work, and preparation of the draft manuscript.

Bogdan I., **Rivers J.**, Coca D. and Beynon R. J. (2008) 'High-performance hardware implementation of a parallel data-base search engine for real-time peptide mass fingerprinting' (submitted to Bioinformatics)

I acquired MS data and performed peptide mass fingerprinting. In addition, I supplied MS data for comparative analysis and was involved in discussions with co-authors prior to preparation of manuscript.

**Rivers J.**, McDonald L., Edwards I. J. and Beynon R. J (2008) 'Asparagine deamidation and the role of higher order protein structure' Journal of Proteome Research 7 (3), 921-7

I discovered that deamidation of peptides occurs during proteolysis, shedding light on previous observations of variable degrees of deamidation of the GAPDH N-terminal peptide (I.J. Edwards & L. McDonald). I was subsequently responsible for all major experimental work to investigate the two processes of proteolysis and subsequent deamidation, and preparation of the draft manuscript. Planning of appropriate experimental work and calculations of reaction kinetics was achieved in discussion with R.J. Beynon.

**Rivers J.**, Simpson D. M., Robertson D. H. L., Gaskell S. J. and Beynon R. J. (2007) 'Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT' Molecular and Cellular Proteomics 6 (8), 1416 – 1427

From initial publication of the QconCAT method (Beynon *et al.*, 2005), I was responsible for implementation of the method for a complete biological study, designed to assess sources of variance and statistical behaviour. Experimental work and preparation of the manuscript was planned in discussion with co-authors, and produced by myself with input from R.J. Beynon.

Bogdan I., Coca D., **Rivers J.** and Beynon R. J. (2007) 'Hardware acceleration of processing of mass spectrometric data for Proteomics' Bioinformatics 23 (6), 724-731

I was responsible for acquiring and supplying MS data, in addition to data processing for comparison of the two methods. For this, I was involved in discussions with co-authors and was responsible for production of several figures, in addition to written sections of the manuscript.

Pratt J. M., Simpson D. M., Doherty M. K., **Rivers J.**, Gaskell S. J. and Beynon R. J. (2006) 'Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes' Nature Protocols 3 (2), 1029 – 1043

I participated in discussions with co-authors and contributed personal experience of using the QconCAT method. In particular, I was responsible for supplying methods of co-digestion of QconCAT with analyte protein, in addition to quantification analysis by MS. I also supplied data for production of relevant figures.

*CATEGORY: Gene Expression*

# High-performance hardware implementation of a parallel database search engine for real-time peptide mass fingerprinting

Istvan Bogdan[1], Jenny Rivers[2], Robert J Beynon[2] and Daniel Coca[1 *]

[1]Dept. of Automatic Control & Systems Engineering, The University of Sheffield, Mappin Street, Sheffield S1 3JD, UK,
[2]Proteomics and Functional Genomics Group, Faculty of Veterinary Science, University of Liverpool, Crown Street, Liverpool L69 7ZJ, UK.

**ABSTRACT**

**Motivation:** Peptide Mass Fingerprinting (PMF) is a method for protein identification in which a protein is fragmented by a defined cleavage protocol (usually proteolysis with trypsin), and the masses of these products constitute a 'fingerprint' that can be searched against theoretical fingerprints of all known proteins. In the first stage of PMF, the raw mass spectrometric data are processed to generate a peptide mass list. In the second stage this protein fingerprint is used to search a database of known proteins for the best protein match. Although current software solutions can typically deliver a match in a relatively short time, a system that can find a match in real-time could change the way in which PMF is deployed and presented. In a paper published earlier (Bogdan *et al.*, 2007) we presented a hardware design of a raw mass spectra processor that, when implemented in FPGA hardware, achieves almost 170-fold speed gain relative to a conventional software implementation running on a dual processor server. In this paper we present a complementary hardware realisation of a parallel database search engine that, when running on a Xilinx Virtex 2 FPGA at 100MHz, delivers 1800-fold speed-up compared with an equivalent C software routine, running on a 3.06GHz Xeon workstation. The inherent scalability of the design means that processing speed can be multiplied by deploying the design on multiple FPGAs. The database search processor and the mass spectra processor, running on a reconfigurable computing platform, provide a complete real time PMF protein identification solution.

## 1 INTRODUCTION

Of the four key enabling tools and technologies in proteomics (protein separation, mass spectrometry instrumentation, protein databases and data processing, analysis and interpretation) it can be argued that the bioinformatics solutions lag in terms of performance and throughput. Post-instrument data processing in particular, is a major bottleneck in the proteomics workflow and as experimental design, instrument performance and user skills increase, it is expected this will become worse. A reasonable goal is that the processing and first-level (non interpretative) analysis should be completed within the same timeframe as the experiment itself, preferably implemented as 'near-instrument' capabilities, where the user has the option of controlling the search parameters and strategies within the timeframe and constraints of the experiment. This largely rules out widely distributed multiprocessor computational farms/grids. However, computational platforms based on a relatively low number of conventional/multicore microprocessors are unlikely to deliver the speed required for real-time processing.

A solution to this problem, advocated in an earlier paper (Bogdan *et al.*, 2007), is to implement proteomics data processing algorithms directly in hardware, effectively designing dedicated digital processors for every algorithm. Central to this approach is a rather unusual digital device termed a Field Programmable Gate Array (FPGA), which allows implementation and operation of a digital hardware design with the same facility as a conventional computer program. Early uses of FPGA devices in biocomputation were to accelerate gene sequence analysis (Fagin et *al.* 1993). FPGAs, which are well suited for high-performance, high-bandwidth and parallel processing applications, have been successfully employed to speed up DNA sequencing algorithms (Hughey 1996, Guerdoux-Jamet *et al.*, 1997, Wozniak 1997, Lavenier, 1998, Guccione *et al.*, 2002, Simmler *et al.*, 2004). FPGAs were also used in the attempt to accelerate search of substrings similar to a template in a proteome (Marongiu *et al.*, 2003). More recently, FPGAs have been used to accelerate sequence database searches with MS/MS-derived query peptides (Anish *et al.*, 2005). This hardware-based solution can locate a query within the human genome about 32 times faster than a software implementation running on a 2.4GHz processor. A hardware sequence alignment tool implemented in FPGA is also available (Oliver *et al.*, 2005). FPGA computing is used for comparing protein sequences with profile HMMs (Sun *et al.*, 2007). An FPGA based BLAST search was used for EST sequencing (Panitz *et al.*, 2007).

In succession to our earlier paper (Bogdan *et al.*, 2007), the focus in the present paper is to use FPGA-based computing to accelerate Peptide Mass Fingerprinting. PMF is a protein identification technique in which a protein is proteolysed using an endopeptidase of defined specificity (usually trypsin) and the masses of the ensuing limit peptide fragments are measured. The proteins are identified by matching the measured molecular masses of these peptide fragments against theoretical peptide mass profiles generated from protein sequence database. PMF is readily delivered at high sensitivity through routine instrumentation such as MALDI-ToF mass spectrometers and although tandem MS approaches can recover more information from single peptides, PMF still plays an important role. Indeed, as more genomes are sequenced, and cross-species matching methods are developed, PMF may assume greater importance for many sub-proteome studies.

PMF involves two basic operations. The first is processing of the raw mass spectrum to derive a mass fingerprint, generating a data set in which the only variable is the mass of each peptide (relative intensities of different ions are not routinely used in PMF). A hardware design of a raw mass spectrum processor that performs this operation was presented previously (Bogdan *et al.*, 2007). When implemented in FPGA hardware, this solution delivered almost 170-fold speed up compared to a conventional software implementation running on a dual processor server.

This paper addresses the second basic operation, which uses the peptide mass fingerprint to search protein databases for possible matches. Typically, a correlation score is computed between the database entries and the

*To whom correspondence should be addressed.

unknown peptide fragment mass list. The matches with the highest score are used to generate a candidate list of most likely proteins to have generated the PMF.

This paper describes the design and hardware implementation of a highly parallel database search engine which, when searching the entire MSDB database, can generate the final candidate protein list in 240ms. The design, which was implemented and run on a Xilinx XC2V8000 FPGA at 100MHz, achieves an average 1800-fold speed gain compared with an equivalent C software implementation, running on a single 3GHz Xeon Server with 4GBytes of memory.

## 2 METHODS

To create the raw data used to evaluate the FPGA implementation, single proteins were diluted with 50mM ammonium bicarbonate and digested with trypsin at a ratio of protein: enzyme of 50:1. Digestion was carried out at 37°C for 24h after which time, 1µl digested material was spotted onto a MALDI target. This was mixed with 1µl α-cyano hydroxycinnamic acid matrix and analyzed using a Micromass M@LDI mass spectrometer (Waters, Manchester, UK) typically over the m/z range 800-4000.

The designs were implemented and tested on a modular and scalable reconfigurable hardware platform, consisting of a FPGA motherboard equipped with a Xilinx Virtex-II XC2V8000 FPGA (8 million gates) and 4Mbytes RAM, communicating with the host PC server *via* a PCI interface. The motherboard has a Xilinx Virtex-II XC2V8000 FPGA, used to implement user designs – in our case the spectrum processor. A second, smaller Xilinx Spartan-II FPGA implements the PCI interface protocols between the server PC and the FPGA system. Communication between these two FPGA devices is at 40MHz on a 32 bit wide data bus. The motherboard can be configured to have up to three additional FPGA modules that can be plugged into dedicated motherboard slots (Figure 1). At present, only one of these three modules has been used to implement the database search. Each FPGA module has one Virtex-II XC2V8000 FPGA device and 1GB of DDR SDRAM that can hold the entire MSDB protein database (currently the encoded MSDB database is stored on a single module and takes about 680MB). Each module is connected with the motherboard FPGA and with the other two modules via a 64 bit, 66MHz local bus. This architecture enables the implementation of parallel searches at FPGA level as well as across modules.

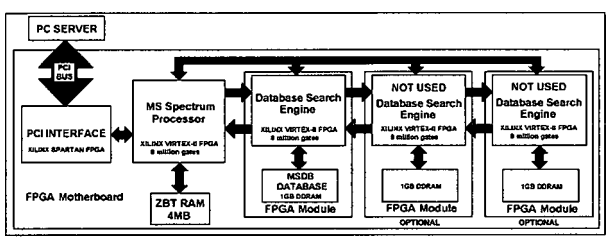



**Figure 1.** Block diagram of the multi-FPGA system. Only one FPGA is used to implement the database search engine.

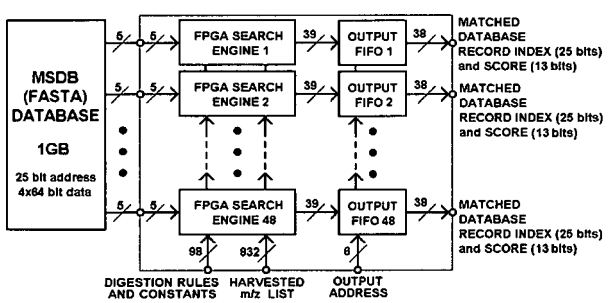



**Figure 2.** Block diagram of the database search engine implemented in FPGA.

The block diagram of the FPGA implemented database search engine is illustrated on Figure 2.

In order to maximize database search speed, the search engine has been configured as a set of 48 identical search processors that can process database records (encoded protein streams) in parallel.

### 2.1 Database Encoding and Storage

In order to reduce communication overheads, the entire MSDB database was encoded and stored in the local 1 GB DDR SDRAM memory available on the FPGA module such that searching the database involves only on-board, local memory access (Figure 1).

The MSDB database is available as plain ASCII text file. In practice, only 20 unique symbols are needed to encode all aminoacids together with some additional standard symbols adopted in the FASTA format. Two additional symbols are used to indicate the end of a protein sequence and the end of the database. In total there are 28 symbols. These can be coded using only 5 bits ($2^5$=32 codes) instead of the minimum 8 required by the ASCII standard, which means that the effective size of the encoded database is about 40% smaller than the original.

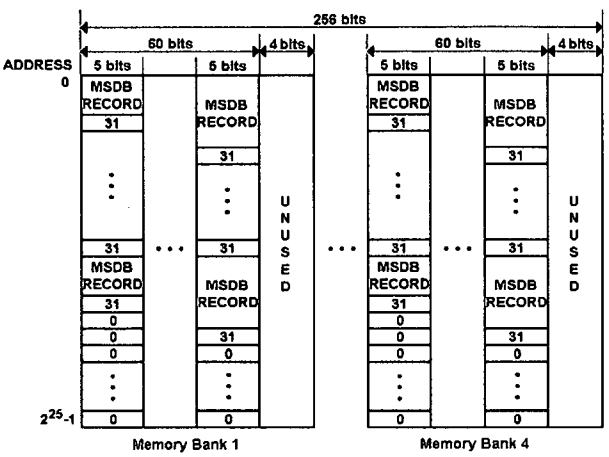



**Figure 3.** Storage of the coded database.

The 1GB DDR SDRAM memory available on the FPGA is organized in 4 banks with a 64-bit wide data bus each. The total data bus width is 256 bits.

Each search engine has a 5-bit wide input, to receive a code-letter at each clock cycle from the database. Each memory module is able to supply 12 data streams of 5 bits synchronously to the corresponding 12 search engines that connect to the output of that RAM bank. The method allows division of the database into 4x12=48 data streams of consecutive records for parallel processing. Each data stream contains a variable number of complete protein sequences, the unused memory locations, which could not hold an entire protein sequence have been padded with zeroes (Figure 3).

### 2.2. Database Search Processor

Each FPGA search processor performs the following basic operations

- *In silico* protein digestion and peptide mass calculation according to externally specified digestion rules and post-translational modifications (currently only fixed modifications implemented).
- Matching score calculation, based on pre-specified mass tolerance

Search results consist of the indexes of the identified species in the database, the matched masses and the total score.

The block diagram of the database search processor is presented in Figure 4. Each search engine is connected to a 5-bit data stream. It reads one code every clock cycle from the corresponding memory column and passes

it to the digestion unit. The digestion unit is responsible for calculating the peptide masses according to the specified digestion rule/parameter. The digestion unit calculates the cumulative mass of the aminoacids received until it encounters a cleavage site, protein record delimiter or the end of database marker. The masses of individual aminoacids, used to compute the peptide masses, are stored into a look-up table as 32 bits fixed-point numbers. When calculating the peptide mass, additional PTM rules can be taken into account.

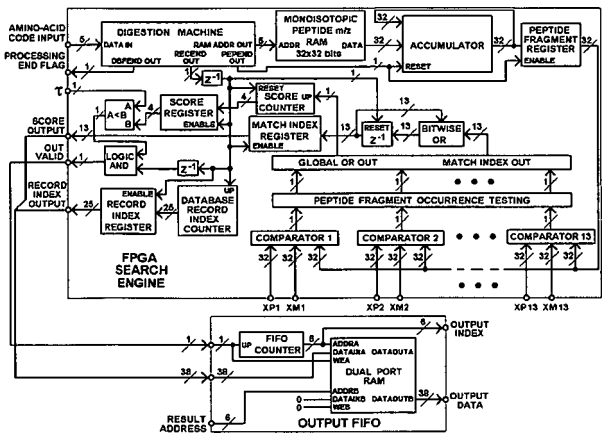
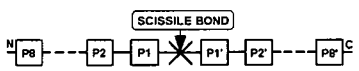



**Figure 4.** Database search processor block diagram.

Figure 5 shows examples of cleavage site rules. The aminoacid residues of the N-terminal side of the scissile bond are noted P1, P2, ..., P8 and the residues of the C-terminal side are noted as P1', P2', ..., P8' according to the Schechter and Berger nomenclature for the description of the protease subsites (Schechter *et al.* 1967). Cleavage site rules are given according to this notation.




EXAMPLES OF CLEAVAGE SITE RULES

| Enzyme | P1 | P1' | P2' |
|---|---|---|---|
| Arg-C | R | X | X |
| Asp-N | X | D | X |
| Chymotrypsin 1 | F, Y, W | X | X |
| Chymotrypsin 2 | A,L,M,F,Y,W | X | X |
| Clostripain | R | X | X |
| Glu-C 1 | E | X | X |
| Glu-C 2 | D, E | X | X |
| Lys-C | K | X | X |
| Trypsin | R, K | X not P | X |
| Thermolysin | X | A,I,L,M,F,V | X not P |
| CNBr | M | X | X |
| Formic acid | D | P | X |

**Figure 5.** Examples of cleavage site rules

The scoring unit counts the matches between the peptide mass fingerprint supplied externally and the peptide masses streamed out from the digestion unit. The precision of the comparison between the m/z values to be searched ($X_1$, $X_2$, ..., $X_n$) and the in-silico computed peptide fragments m/z values ($Y_i$) can be specified by the user-defined parameter $R$ that holds the desired error tolerance (as ppm) which may be dictated by the MS instrument accuracy. A match is found if

$$XM_i < Y_i < XP_i, i = 1,\ldots,n$$

$$XM_i = X_i\left(1-\frac{2R}{10^6}\right)\Big/\left(1-\frac{R}{10^6}\right), XP_i = X_i\Big/\left(1-\frac{R}{10^6}\right)$$

When a valid peptide fragment mass $Y_i$ is computed, it is compared in parallel to peptide mass fingerprint $X_1, X_2, ..., X_n$). In the current implementation $n$=13 so the computed peptide mass $Y_i$ is compared in parallel with 13 m/z values. The result of the comparison is used to generate the basic cumulative score for every processed protein.

The number of m/z values that are used in the search can however be increased at the expense of increasing the complexity of the design of individual search processors. This normally means that the number of processors that can be allocated on a single FPGA (currently 48) may need to be reduced. Since the search can be easily deployed on multiple FPGA modules, performance can be preserved or even expanded by adding additional modules. Alternatively, the user can decide the best tradeoff between performance and search strategy.

If a match is found, the score counter is incremented by one. The position of a match is also recorded in an $n$-bit ($n$=13 here) match index word. When the end of a record is found, the record index counter, the score counter and the match index register outputs are stored in intermediate registers.

If a peptide fragment has multiple occurrences in a database record, and it is found to match one of the m/z peaks from the input peptide mass fingerprint list, the score counter is incremented only once, after the first occurrence. A peptide fragment occurrence block is responsible for this function.

Each search processor has 3 outputs: a processing end flag that remains set after processing ends until the search processor is reset; an output index that remains set to the last available FIFO address where the total the number of matches is stored and a 39 bits output that contains the results (database index.

Results of the 48 search engines are collected in dual port RAM devices organized as FIFO structures of 64 words of 39 bits each. The user can specify a score threshold $\tau$ so that only the matches that are above a threshold are saved i.e. if the score of a given match is higher than a programmable threshold $\tau$, the corresponding record index and match index are stored in the output FIFO.

The basic matching score can be used to implement more sensitive scoring schemes which account for peptide frequency distributions such as MOWSE (Pappin *et. al.* 1993), PIUMS (Samuelsson *et al.* 2004) or more comprehensive Bayesian scoring approaches which also account for the individual properties of the proteins analyzed such as ProFound (Zhang and Chait, 2000). These scoring schemes are have not yet been implemented in FPGA. Currently these scoring algorithms are run on the PC server. The externalization of the scoring statistics means that the output of the search can be rapidly evaluated using different scores, and even developed into a consensus score validation scheme.

### 2.3. FPGA implementation

The actual database search design occupies about 99% of the FPGA's logic resources, 99% of the FPGA's internal RAM resources and 53% of the FPGA's I/O resources. The 1GB on-board RAM is capable of a data rate of 100MHz/cycle which is the processing speed of the implemented database search engines. However the data transfers between the separate FPGA devices are at slower speed of 40MHz. The design includes all necessary control and FIFO structures that implement a 64-bit wide data transfer between the FPGA devices. The FPGA board was installed and tested on Single & Dual 3.06 GHz Xeon processor servers with 4GB RAM under Windows XP Professional.

All arithmetic operations on the m/z values were performed using 32-bit unsigned fixed-point binary number representation of mass and abundance values, with 12 bits after the radix point.

## 3 RESULTS

### 3.1 Speed gains

The MSDB database (31/08/2006) was encoded and loaded in the local 1GB DDR SDRAM module memory. The database contains 3,239,079 records with 1,079,594,700 effective code letters. If the additional separator

codes are included the encoded database requires 1,082,833,779 symbols and occupies 67% of the available 1GB.

A reference C program models the exact computational flow implemented by the hardware design. In tests, the output results of both the software and FPGA implementation of the database search engine are identical. The C program was run on Single and Dual 3.06GHz Xeon PC servers under WindowsXP Professional. In all simulations the FPGA implementation matched correctly the simulated peptide mass fingerprints

The performance of both software and hardware (FPGA) designs were assessed using a randomly selected database records that were digested *in-silico* using trypsin digestion rules. In each case, the search was carried out using 13 peptide m/z values selected randomly from the theoretical protein digests. The processing time for the software implementation accounts only for the main computational loop, after all variables have been initialized.

The FPGA system performs a complete database search in 240 (±0.02)ms while the completed average processing time for the C implementation is 7.2min. As seen in Figure 6, the speed gain of the FPGA over the C software implementation ranges from 1650-1950 fold average speed-up.

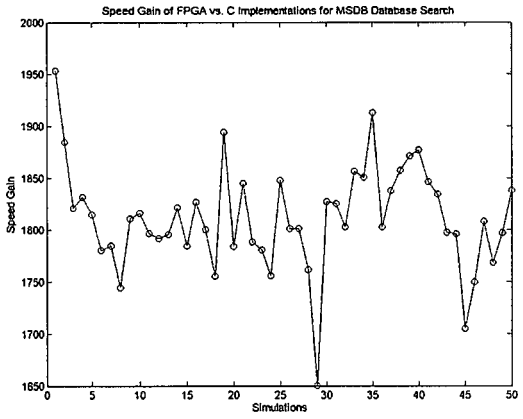



**Figure 6.** Speed gain of FPGA vs. C implementation of MSDB database search

The average processing time of the C implementation on a dual processor server is 6.8min. The average speed gain in this case is 1695. It is worth noting that the search time quoted above could be reduced by pre-indexing the protein database – this strategy is employed by commercial software solutions. Processing time on FPGA is not dependent of the peptide mass fingerprint data set and has linear dependency with the database size. To illustrate this aspect we have performed FPGA searches using only the database of *Saccharomyces cerevisiae* proteins. This database occupies only 0.277% of the RAM allocated for the full MSDB. In this case, the search time decreased to 0.66ms which represents 0.275% of the full database search time. In practice various subset databases can be loaded on-the fly the memory of the FPGA search module to carry out significantly faster single-species PMF searches, for example.

### 3.2 Experimental validation

We have tested the matching accuracy of the FPGA system using peptide mass fingerprints generated from raw MALDI-TOF mass spectra. Raw MALDI-TOF data were processed using the mass spectra processor detailed in Bogdan *et al.* (2007). The PMF consisted of m/z values selected from the identified peak list after the elimination of known contaminants such as trypsin and keratin.

In each case, the search was carried out using the FPGA system implementation and MASCOT (www.matrixscience.com) with identical search parameters. In all cases, the same peptide matches were returned by both systems. For illustration, Table 1 shows the matching results from a batch of ten mass spectra. Figure 7 shows examples of raw and processed mass spectra for actin (chicken cardiac muscle). Table 2 shows the list of m/z values used in this particular search.

| Protein | No Matches – FPGA/Mascot |
|---|---|
| A23022 actin, cardiac muscle - chicken | 6 |
| 1HBRB chicken hemoglobin, beta chain - chicken | 5 |
| KICHPM pyruvate kinase, muscle - chicken | 8 |
| G3P_CHICK Glyceraldehyde-3-phosphate dehydrogenase (GAPDH).- chicken. | 6 |
| S12151 L-lactate dehydrogenase (EC 1.1.1.27) chain A - chicken | 9 |
| PGAM1_CHICK Phosphoglycerate mutase 1 chicken. | 4 |
| ISCHT triose-phosphate isomerase - chicken | 7 |
| KICHCM creatine kinase chain M - chicken | 7 |
| HACH1 hemoglobin alpha chain - chicken | 5 |

**Table 1.** Peptide matches returned by FPGA/Mascot searches for a set of experimental MS data.

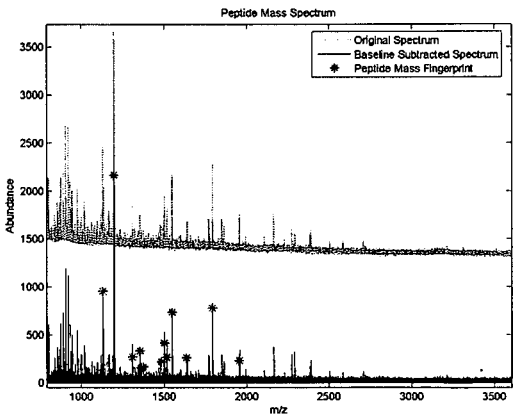



**Figure 7.** Example of raw and processed mass spectrum - actin. The PMF used to search the database is indicated by *.

## 4 DISCUSSION

We have successfully demonstrated that an MSDB PMF database search can be implemented in FPGA. If the FPGA database search module is integrated with the FPGA spectrum processor described in (Bogdan *et al.*, 2007) the spectrum processing and database search time is around 241ms. This means that the complete hardware PMF solution can process and match 4 mass spectra per second. This figure can be enhanced if the database search is distributed across multiple FPGAs and even with the current board it should be possible to achieve processing speeds of 12 spectra per second. Further expansion is possible if more than one FPGA motherboards are configured. The current server could hold up to 3 FPGA boards so in principle such a reconfigurable computing node can deliver up to 36 PMF protein matches per second.

| PMF List (m/z) | Matched peptide masses by FPGA & Mascot | Error [ppm] |
|---|---|---|
| 1130.63 | 1130.54 | 79.6 |
| 1198.80 | 1198.70 | 83.4 |
| 1307.76 | | |
| 1354.74 | | |
| 1364.83 | | |
| 1374.86 | | |
| 1483.83 | | |
| 1500.81 | 1500.70 | 73.3 |
| 1515.84 | 1515.74 | 66.0 |
| 1547.87 | | |
| 1634.90 | | |
| 1791.03 | 1790.89 | 78.2 |
| 1956.14 | 1956.04 | 51.2 |

**Table 2.** Example of experimental PMF (actin) and the corresponding matched theroretical peptide masses. Tolerance was set at 250ppm.

As new genome sequences are completed, the size of the corresponding proteome databases will also grow. In our tests, the whole encoded database was loaded into the local memory available on the FPGA search module. If more than one search modules are included, the database can be split and distributed across different search modules. Alternatively, it is rare for the experimenter not to know the species with which they are working, and a degree of taxonomic restriction to generate a logically reduced database is entirely feasible. Encoded databases can be switched in and out of FPGA RAM very rapidly (~500ms for the full MSDB). Since the search speed increases linearly with database size, sub-setting the database is an efficient way of increasing search speed.

At present the implementation does not include the flexibility to match variable post-translational modifications (www.unimod.org). There are several strategies that might be employed to permit such modifications, either by extension of the amino acid types (unmodified or modified) or by extension of the database entries to include fixed and variable modifications. However, the performance of the current, minimally specified system is a powerful advocacy for extension of the software to include modification searching. Additionally, we have deliberately externalized to the FPGA the scoring algorithm, as this does not explicitly influence the search itself and therefore provides more flexibility for invocation of alternative scoring methods (Pium [Samuelsson *et al.* 2004], Peptident [Gattiker *et al.* 2002], MASCOT [Perkins *et al.* 1999]), and the possibility of consensus scoring approaches. Further work is underway to further optimize and refine the designs and to provide additional features and functionality.

## ACKNOWLEDGEMENTS

This work was funded by BBSRC (Grant No. BBS/B/16402). The authors gratefully acknowledge the support of Xilinx Inc. who donated the devices and design tools used in this study.

## REFERENCES

Anish T. A., Dumontier M., Rose J. S., Hogue C. W. V. (2005) Hardware-accelerated protein identification for mass spectrometry. *Rapid Communications in Mass Spectrometry*, **19**, 833-837.

Bogdan I., Coca D., Rivers J., Beynon J. R. (2007) Hardware acceleration of processing of mass spectrometric data for proteomics. *Bioinformatics Gene Expression*, **23**, 724-731.

Fagin B., Watt J. G., Gross R. (1993) A special-purpose processor for gene sequence analysis. *Comput. Appl. BioSci.*, **9**, 221-226.

Gattiker A., Bienvenut W. V., Bairoch A., Gasteiger E. (2002) FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomix*, **2**, 1435-1444.

Guccione A. S., Keller E. (2002) Gene Matching Using Jbits, Proceedings of the Reconfigurable Computing Is Going Mainstream, 12th International Conference on Field-Programmable Logic and Applications, 1168 – 1171.

Guerdoux-Jamet P., Lavenier D. (1997) SAMBA: hardware accelerator for biological sequence comparison, *Comput. Appl. BioSci.*, **13**, 609-615.

Hughey R. (1996) Parallel hardware for sequence comparison and alignment. *Comput. Appl. BioSci.*, **12**, 473-479.

Lavenier D. (1998) Speeding up genome computations with systolic accelerator", *SIAM News*, **31**, 1-8.

Marongiu A., Palazzari P., Rosato V. (2003) Designing hardware for protein sequence analysis. *Bioinformatics*, **19**, 1739-1740.

Oliver T., Smidth B., Nathan D., Clemens R., Maskell D. (2005) Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics*, **21**, 3431-3432.

Panitz F., Stengaard H., Hornshoj H., Gorodkin J., Hedegaard J., Cirera S., Thomsen B., Madsen L. B., Hoj A., Vingborg R. K., Zahn B., Wang X., Xuefey W., Wernersson R., Jorgensen C. B., Scheibye-Knudsen K., Troels A., Lumholdt S., Sawera M., Green T., Nielsen B. J.,Havgaard J. H., Brunak S., Fredholm M., Bendixen C. (2007) SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation. *Bioinformatics*, **23**, i387-i391.

Pappin D. J. C., Hojrup P., Bleasby A. J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, 3, 327-332.

Schechter I., Berger A. (1967) On the size of the active site in proteases. *Biochem. Biophys. Res. Com.*, **27**, 157-162.

Perkins D. N., Pappin D. J. C., Creasy D. M., Cottrell J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Samuelsson J., Dalevi D., Levander F., Rögnvaldsson T. (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting, *Bioinformatics*, **20**, 3628-363.

Simmler H., Singpiel H., Männer R. (2004) Real-Time Primer Design for DNA Chips. *Interscience Concurrency Computation: Practice and Experience*, **16**, 855–872.

Sun Y., Buhler J. (2007) Designing patterns for profile HMM search. *Bioinformatics*, **23**, e36–e43.

Zhang, W. and Chait, B.T., (2000). ProFound: An Expert System for Protein Identification Using Mass Spectrometric Peptide Mapping Information. *Anal. Chem.*, vol. 72, 2482-2489.

research articles Journal of proteome research

# Asparagine Deamidation and the Role of Higher Order Protein Structure

**Jenny Rivers, Lucy McDonald, Ian J. Edwards, and Robert J. Beynon***

*Proteomics and Functional Genomics Group, Faculty of Veterinary Science, University of Liverpool, Crown Street, Liverpool L69 7ZJ, United Kingdom*

**Received July 11, 2007**

The 'protein world' exhibits additional complexity caused by post-translational modifications. One such process is nonenzymic deamidation of asparagine which is controlled partly by primary sequence, but also higher order protein structure. We have studied the deamidation of an N-terminal peptide in muscle glyceraldehyde 3-phosphate dehydrogenase to relate three-dimensional structure, proteolysis, and deamidation. This work has significant consequences for identification of proteins using peptide mass fingerprinting.

**Keywords:** Deamidation • proteolysis • protein structure • asparagine • aspartic acid • peptide mass fingerprinting

## Introduction

The emergence of new analytical methods for protein characterization has led to the recognition that there is an additional dimension of complexity in the protein world created by a wide range of post-translational modifications. Some of these modifications are specific and are part of the obligatory maturation process of a protein, such as the removal of propeptides. Other changes are transient, reversible, and may only operate on a subset of molecules in the protein pool (the best understood is phosphorylation). Other irreversible changes, such as deamidation or lysine aldehyde mediated cross-linking, are nonenzymic, and the longevity of the protein may be reflected in the accumulation of such changes.

Deamidation of the side chain of asparagine residues is a nonenzymic process[1] (www.deamidation.org). The conversion of asparagine to aspartic acid or isoaspartic acid elicits a local change in charge, and has the potential to impose a self-timer on protein molecules, altering activity or stability with lifetime kinetics.[2-5] The ability to include a nonenzymic irreversible change into a protein that elicits a small steric change but a substantial local alteration in electrostatic potential could provide an opportunity to evolve a programmable irreversible change of state into a protein. Most studies on asparagine deamidation have been conducted with model peptides[6] which are essentially devoid of higher order structure and which permit the peptide backbone and side chain to adopt a conformation compatible with the cyclic intermediate that is required for this reaction to take place. Since the flexibility and conformational freedom of the peptide is modified by the nature of the amino acids, the rate of deamidation of model peptides is strongly influenced by the flanking residues[6] and the primary influence on the rate of asparagine deamidation is the amino acid C-terminal to the asparagine residue. From studies of model peptides, the highest rate of deamidation is obtained when the carboxyl neighbor is glycine, yielding a half-time for deamidation of around 24 h.[6] This is probably because the lack of a C$\beta$ atom minimizes steric hindrance and permits ready formation of the five-membered imide conducive to the deamidation reaction. The N-terminal neighbor has a minor effect on the rate of deamidation.[6] While most of our understanding of rates of peptide deamidation has derived from short, model peptides, the same sequences, when incorporated into protein structures, might acquire a relatively immobile backbone trajectory that could constrain the sequence to either favor or disfavor deamidation.

The resolution of modern mass spectrometers used routinely in proteomic analyses permits ready resolution of the monoisotopic peptide–ion from the $^{13}C$ isotopomer variants, even at charge states of +2 or +3. At this level of resolution, a deamidation event (Asn → Asp) would be readily recognized, as it elicits a mass shift of +0.985 Da (-$NH_2$ = 16.03 to -OH = 17.01). In circumstances where a peptide exists as a mixture of the amide and cognate acid species, a complex mass spectrum would ensue that appears as an atypical isotopomer distribution for a peptide of that mass. It follows that partial deamidation events should be readily observed by examination of the atypical profile, particularly without prior chromatographic separation that would resolve the amide and cognate acid in chromatographic space.

In the course of proteomics studies of soluble proteins in skeletal muscle,[7] we observed that a peptide from one protein in particular exhibited a noticeable and atypical natural isotope distribution profile, consistent with a mixture of an asparagine-containing peptide and the cognate deamidation product. This peptide was derived from the N-terminus of an abundant protein, glyceraldehyde-3-phosphate dehydrogenase (GAPDH). We present here a comprehensive analysis that confirms that the 'atypical' isotope profile is in fact attributable to partial deamidation of an asparagine residue. Deamidation of -AsnGly-

* To whom correspondence should be addressed. Phone: +44 151 794 4312. Fax:+44 151 794 4243. E-mail: r.beynon@liv.ac.uk.

10.1021/pr070425l CCC: $40.75 


Published on Web 02/05/2008

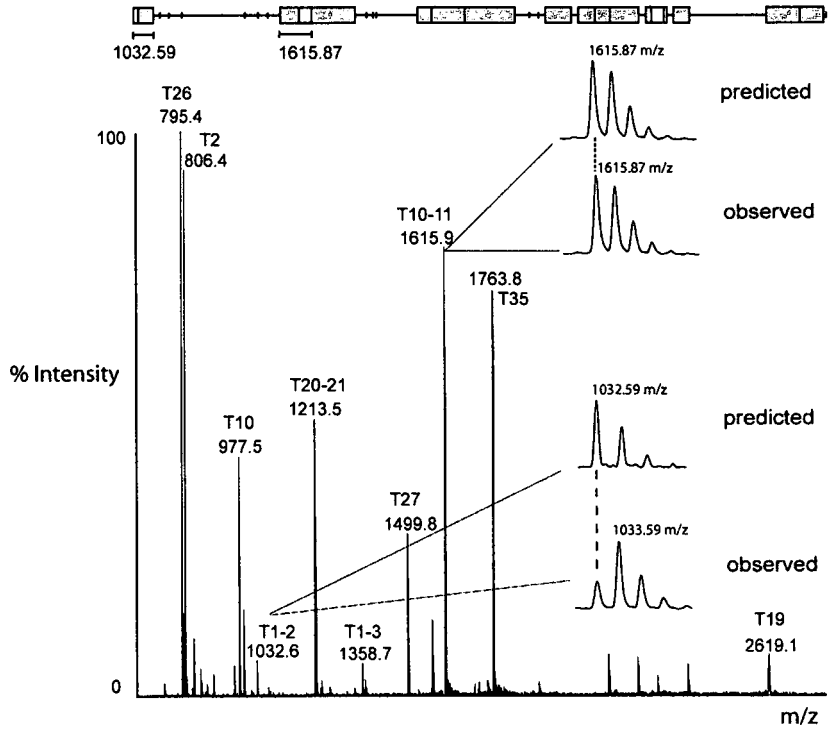



**Figure 1.** Atypical peptide mass spectrum consistent with deamidation. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH; 1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) purified from rabbit skeletal muscle (Sigma, Dorset, U.K.) was digested in solution with trypsin at a substrate/protease ratio of 100:1 by weight, and the masses of the resultant tryptic peptides were assessed by MALDI-ToF mass spectrometry; a coverage map is included at the top of the figure, with identified peptides indicated by a shaded block and those identified as part of a missed cleavage by an open block. The spectrum of a typical partial cleavage tryptic peptide (T10–11, *m/z* 1615.9) was compared with the mass spectrum predicted by the MS-Isotope tool (http://prospector.ucsf.edu/). This behavior, common to almost all other peptides, emphasized the atypical profile observed for the N-terminal partial cleavage peptide (T1–2, *m/z* 1032.6).

sequences occurs during sample preparation in proteomics,[8] and proteolysis conducted at lower pH and temperature will minimize artifactual deamidation.[9] Here, we show that deamidation is constrained by higher order structure and is enhanced after release of that conformational restraint by proteolysis. This observation has significance for the identification of deamidation events by protein or peptide mass spectrometry[10–12] and reinforces the role that protein conformation can play in this process.

## Experimental Section

**Materials and Reagents.** Trypsin (sequence grade) was obtained from Roche Diagnostics (Lewes, U.K.). All other chemicals and solvents (HPLC grade) were purchased from Sigma-Aldrich Company Ltd. (Dorset, U.K.) and VWR International Laboratory Supplies (Leicestershire, U.K.).

**One-Dimensional Gel Electrophoresis (1DGE).** Purified GAPDH from rabbit skeletal muscle (Sigma, Dorset, U.K.) (10 $\mu$g) was electrophoresed through a 12.5% polyacrylamide gel and visualized with Biosafe Coomassie Brilliant Blue stain (Bio-Rad, Hemel Hempstead, U.K.). Gels were destained with a 10% acetic acid 10% methanol solution.

**In-Gel Trypsin Digestion.** Gel plugs containing GAPDH (identification confirmed by MALDI-ToF MS, results not shown) were excised from 1D gels using a glass pipet and transferred to an Eppendorf tube. To each tube, 25 $\mu$L of 50 mM ammonium bicarbonate, pH 8.2, and 50% (v/v) acetonitrile (ACN) was added and incubated at 37 °C for 20 min. This process was repeated until all of the stain had been removed. The plugs were then washed in 50 mM ammonium bicarbonate, which was subsequently discarded. The gel was dehydrated using 5 $\mu$L of ACN, and incubation at 37 °C was resumed for 30 min. Once dry, the gel was rehydrated in 50 mM ammonium bicarbonate (9 $\mu$L) containing trypsin (1$\mu$L of 100 ng/$\mu$L trypsin stock reconstituted in 50 mM acetic acid), and digestion was allowed to continue overnight at 37 °C; the digestion was halted by the addition of 2 $\mu$L of formic acid.

**MALDI-ToF Mass Spectrometry.** Peptides were analyzed by MALDI-ToF (M@LDI; Waters, Manchester, U.K.) mass spectrometry. For this, 1 $\mu$L of digested material was mixed with an equal volume of α-cyano-hydroxycinnamic acid in 50% (v/v) ACN and 0.1% (v/v) trifluoroacetic acid. This was allowed to dry, and peptides were acquired over the range 900–3000 *m/z*. For each combined spectrum, 20–30 spectra were acquired (laser energy typically 30%) with 10 shots per spectrum and a laser firing rate of 5 Hz. Data were processed using MassLynx software to subtract background noise using polynomial order 10 with 40% of the data points below this polynomial and a tolerance of 0.01. Spectral data were also smoothed by performing two mean smooth operations with a window of three channels. To confirm the assumption that both acid and amide forms of the peptide ionize with equal signal response in MALDI-ToF MS, the synthetic peptide for the amide form was allowed to fully deamidate (by incubation at 37 °C) and mixed in a known ratio with asparagine-containing peptide in a strong acidic solution to prevent further deamidation. The signal response from the two variants was identical.

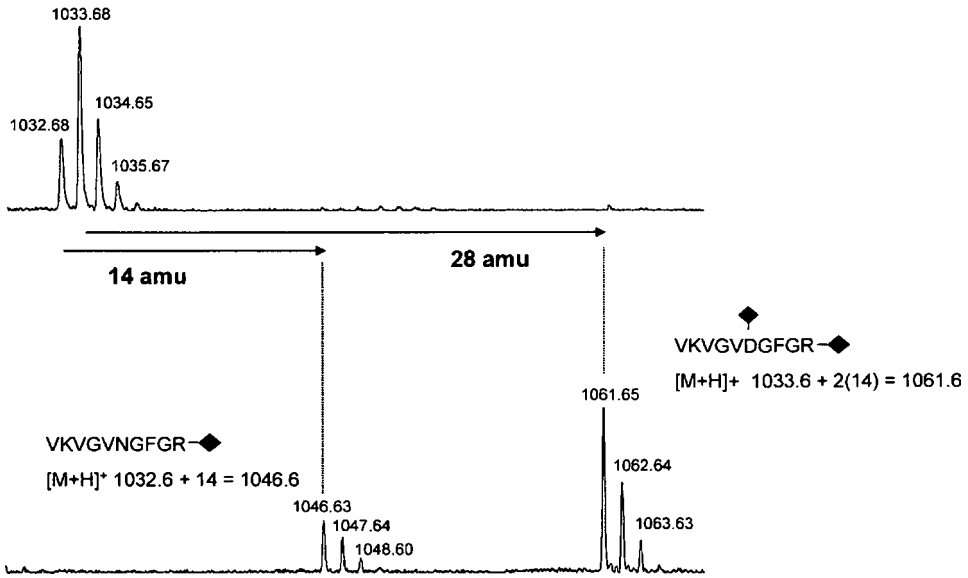



**Figure 2.** Esterification of acidic residues in the N-terminal peptide of GAPDH. Tryptic peptides recovered from an in-gel tryptic digest of GAPDH (purified from rabbit skeletal muscle, Sigma, Dorset, U.K.) were reacted with acetyl chloride and methanol to convert acidic residues to their corresponding methyl esters. The upper mass spectrum is the peptide resulting from partial deamidation of $Asn_6$, thus, is a mixture of two forms (asparagine containing and aspartic acid containing). The lower spectrum, obtained after esterification has resolved the peptide into two distinct reaction products at 1046.63 *m/z* and 1061.65 *m/z*, consistent with the addition of one and two methyl groups (+14.03 Da), respectively.

**In-Solution Tryptic Digestion.** Soluble protein (purified GAPDH; 1 mg/mL) was diluted 10-fold with 50 mM ammonium bicarbonate prior to addition of trypsin (100:1 substrate/protease). The reaction mixture was incubated at 37 °C for 24 h, and peptides were analyzed by MALDI-ToF MS.

**Esterification of Peptides.** A stock solution of methanol (1 mL, previously stored at −20 °C for 15 min) and acetyl chloride (150 $\mu$L) was prepared. An aliquot (10 $\mu$L) of this mixture was then added to a dried portion of the peptide pool recovered after in-gel digestion of the protein. The mixture was incubated at room temperature for 45 min prior to drying in a vacuum centrifuge. Esterified peptides were analyzed by MALDI-ToF MS.

**Monitored Proteolysis of GAPDH.** Digestion reaction mixtures with trypsin were stopped at selected time points after addition of enzyme by removing 10 $\mu$L and adding to an equal volume of 10% (v/v) formic acid. The fractions were subsequently stored at −20 °C until the end of the time course. Peptides were analyzed by MALDI-ToF MS.

**Data Processing.** The natural isotope profile for the acid VKVGVDGFGR and amide VKVGVNGFGR variants of the GAPDH N-terminal peptide were predicted using the MSIsotope tool provided online within the Protein Prospector Package (http://prospector.ucsf.edu/ucsfhtml4.0/msiso.htm). The intensities of each isotopomer peak were added, and the combined theoretical spectrum was compared with the intensities derived from the experimental mass spectrum. The sum of the squares of the deviation between predicted and experimental data was used to generate the object function, and the sole parameter ($P_A$) was the proportion of the acidic component (by definition, equal to $1 - P_N$, where $P_N$ is the proportion of amide). The nonlinear optimization function (Solver) within Excel was used to obtain the best fit value of $P_A$. Additionally, some samples were analyzed by a high speed spectrum deconvolution tool, implemented as computer hardware in a field programmable gate array.[13]

**Absolute Quantification of Proteolysis Using a Stable Isotope-Labeled Synthetic Peptide.** The N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and neutral mass 1041.59 Da, was synthesized by Sigma-Genosys (Dorset, U.K.) and was labeled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset from the analyte peptide. For quantification of proteolysis, the synthetic peptide was added to digested material in 10% (v/v) formic acid to stop digestion and deamidation. Peptides were analyzed by MALDI-ToF MS, and the relative intensities of analyte peptide and internal standard were used to quantify the amount of peptide released from the protein during incubation with trypsin at 37 °C. As conversion of asparagine to aspartic acid alters the isotope envelope of the analyte peptide, the composite abundance of the entire isotopic envelope for both analyte and internal standard peptide was summed in each case. These data permitted the kinetics of proteolytic release of the N-terminal peptide from GAPDH to be calculated and were used along with the kinetics of deamidation to investigate the interaction between these two alternative processes. The rate of deamidation was measured across the time course of digestion by calculating the proportion of acid and amide variants of the peptide at each time point. This was done during proteolysis of GAPDH and for the synthetic peptide, at different temperatures.

## Results and Discussion

One of the most abundant soluble sarcoplasmic proteins in skeletal muscle is glyceraldehyde 3-phosphate dehydrogenase, amounting to 11 ± 1% (mean ± SEM, $n = 3$) of soluble protein when resolved by 1D gel electrophoresis (1DGE) and analyzed by densitometry (data not shown) and up to 500 ± 50 nmol/g (mean ± SEM, $n = 4$) tissue when analyzed using the QconCAT method for absolute quantification.[13–15] MALDI-ToF spectra

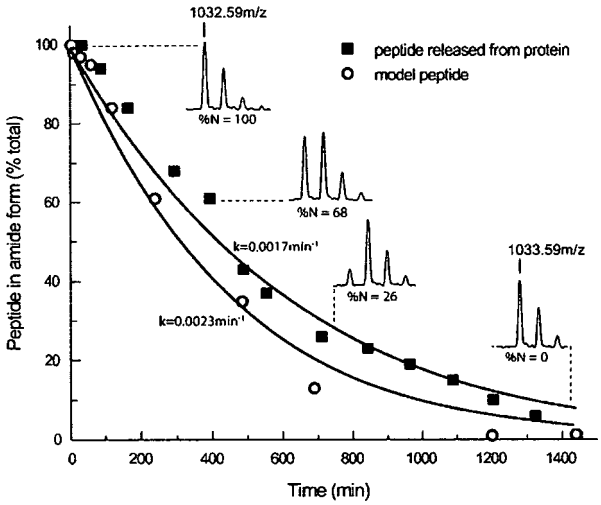



**Figure 3.** Time course of deamidation of the N-terminal peptide of GAPDH. Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, U.K.; 1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) was digested with trypsin (trypsin/protein 1:100) over 24 h at 37 °C. Proteolysis was stopped at 0, 2, 5, 10, 30, 60, 120, 240, 480, and 1440 min by mixing 10 $\mu$L from the digestion mixture with 10 $\mu$L of 10% (v/v) formic acid. The resulting peptides were analysed by MALDI-ToF mass spectrometry, and deamidation was monitored during proteolysis for the N-terminal peptide of sequence VKVGVNGFGR at 1032.59 *m/z*. The proportion of acid and amide variants was assessed as described in Experimental Section, from peak height data, and plotted as a function of time (closed squares). Peptide envelopes illustrating the conversion of acid to amide form in MALDI-ToF mass spectra corresponding to time points over 24 h are inserted above the data. To compare this with model peptide studies, the N-terminal peptide of GAPDH, of sequence VKVGVNGFGR and mass 1041.59 Da, was synthesised by Sigma-Genosys (Dorset, U.K.) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset from the analyte peptide. This peptide was incubated in 50 mM ammonium bicarbonate at 37 °C, and a sample of the peptide was added to an equal volume of 10% (v/v) formic acid at selected time points. The relative amounts of acid and amide variants of the peptide were measured using MALDI-ToF MS, and this was used to calculate the rate of deamidation. These data are presented as open circles. The solid lines are the trajectories taken by first-order decay for the synthetic peptide and the proteolyzed glyceraldehyde 3-phosphate dehydrogenase.

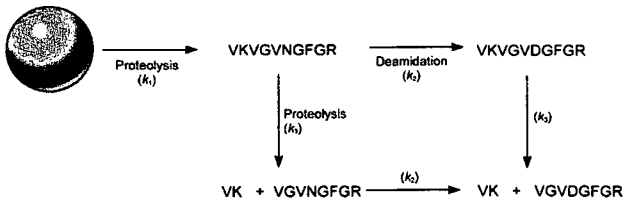



**Figure 4.** Model of proteolysis and deamidation of the N-terminal peptide of GAPDH. The simultaneous processes of proteolysis and release of the N-terminal peptide of GAPDH followed by deamidation of the asparagine residue to aspartic acid were modelled according to this scheme. The model also included the subsequent proteolysis of the N-terminal peptide (**VKVGVNGFGR** or **VKVGVDGFGR**) at the internal arginine residue to generate a dipeptide and a truncated peptide (**VK**+**VGVNGFGR** or **VK**+**VGVDGFGR**). In this scheme, we assumed that the rate of deamidation was the same, whether in the full length or truncated N-terminal peptide, and that the rate of removal of the N-terminal dipeptide was independent of the amide/acid variants.

for this protein, isolated by 1DGE and digested with trypsin prior to MS analysis are of high quality, give very high probability identification of this protein (not shown), and yield approximately 20 peptides, ranging from 805.5 *m/z* to 2265.4 *m/z*. Close inspection of each peptide indicated that for most, the observed mass isotopomer distribution was as expected, and was in close agreement to the distribution predicted by the MsIsotope program (http://prospector.ucsf.edu/). One peptide in particular (VKVGVNGFGR, $[M+H]^+$ 1032.58 *m/z*) was notably different from the others, inasmuch as the isotope distribution profile was far removed from the predicted profile (Figure 1). In particular, the relative intensity of the monoisotopic ion was diminished, and of lower intensity than the first [$^{13}C$] isotopomer, a relative intensity pattern that is unexpected for a peptide of mass 1031.58 Da, given an empirical formula of $C_{46}H_{78}N_{15}O_{12}$.

The mass isotopomer envelope is consistent with the analyte being a mixture of two peptides, one of monoisotopic *m/z* 1032.58 and a second at a monoisotopic *m/z* of 1033.58. The higher *m/z* peptide could have been a contaminant or it could have been generated from the peptide at *m/z* 1032.58. In the latter case, the most probable explanation for the mass increase was deamidation of the asparagine residue, which, by conversion to an aspartate residue, would increase the mass by 0.985 Da (-$NH_2$ to -OH). To prove that the atypical profile was a consequence of deamidation, we esterified the peptide mixture to convert carboxyl groups to their methyl esters. The mass shift on esterification would be 14.03 Da. Because the peptide $\mathbf{V_2KVGVNGFGR_{10}}$ would possess a single carboxyl group in the amide form (the alpha carboxyl group), and two in the acid form, esterification should therefore deconvolute the atypical peptide into two products, one esterified at a single position (+14.03 Da), and a second modified in two positions (+28.06 Da). When the peptide mixture was analyzed after esterification, the MALDI-ToF ions in the 1032−1036 *m/z* region disappeared, and two new ions appeared, one representing the single modified amide (*m/z* 1032.58 + 14.03 = 1046.61) and the second reflecting the double modified acid (*m/z* 1033.58 + 28.06 = 1061.64; Figure 2).

From this analysis, it was not possible to assess whether the residue had deamidated *in vivo* or was an artifact of sample preparation and processing. To assess the extent of deamidation of this peptide in the native protein, we treated purified rabbit GAPDH with trypsin and monitored the proteolysis and the partition between the acid and amide variants of the peptide in MALDI-ToF mass spectra (Figure 3; the same experiments were repeated for an in-solution tryptic digest of chicken skeletal muscle soluble proteins and the same behavior was apparent, results not shown). The N-terminal peptide of GAPDH (**VKVGVNGFGR**) was released within a few minutes and was readily detected as the first analyte ion to appear in the MALDI-ToF spectrum. In the early stages of digestion, the mass spectrum of this peptide was entirely consistent with it being exclusively in the amide form. However, as time progressed during proteolysis, the mass spectrum of the peptide showed that the peptide was converted to a mixture of the amide and acid variants, and after 10 h of digestion, the peptide was over 80% in the acid form. The first-order rate constant for this process was approximately 0.0017 $min^{-1}$, which was higher than the value derived from model peptides; for the sequence $NH_2$**GVNGG**OH, the first-order rate constant was

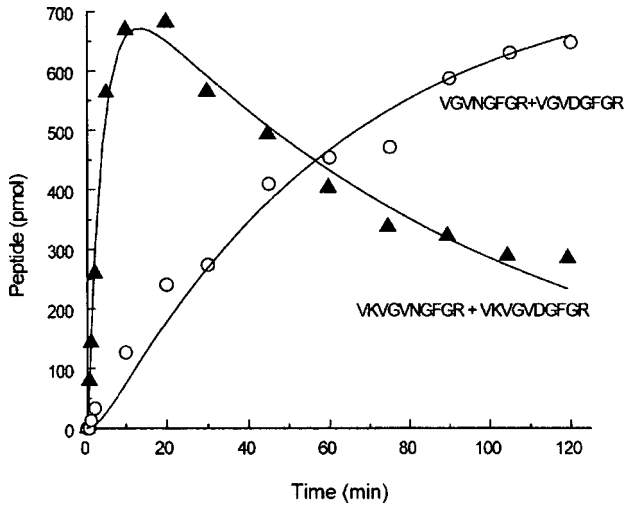



**Figure 5.** Absolute quantification of proteolysis of the GAPDH N-terminus. Purified rabbit skeletal muscle GAPDH (Sigma, Dorset, U.K.; 1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) was digested with trypsin (trypsin/protein 1:10) over 24 h at 37 °C. The N-terminal peptide of GAPDH, of sequence **VKVGVNGFGR** and mass 1041.59 Da, was synthesised by Sigma-Genosys (Dorset, U.K.) and was labelled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset from the analyte peptide. For quantification of proteolysis, the synthetic peptide was added to digested material in 10% (v/v) formic acid to stop digestion at selected time points. Peptides were analyzed by MALDI-ToF MS, and the relative intensities of analyte peptide and internal standard were used to quantify the amount of peptide released from the protein during incubation with trypsin at 37 °C. Both the N-terminal peptide (**VKVGVNGFGR/VKVGVDGFGR**; *m/z* 1032.59 $[M+H]^+$; closed triangles) and the shorter peptide produced by further proteolysis (**VGVNGFGR/VGVDGFGR**; *m/z* 805.59 $[M+H]^+$; open circles) were monitored. As conversion of asparagine to aspartic acid alters the isotope envelope of the analyte peptide, the composite abundance of the entire isotopic envelope for both analyte and internal standard peptide was summed in each case. The solid lines reflect the fitted curves for the transient appearance of the N-terminal peptide (**VKVGVNGFGR/VKVGVDGFGR**) and the truncated product (**VGVNGFGR/VGVDGFGR**), modelled and fitted as sequential first-order reactions (see text).

previously measured at 0.0004 $min^{-1}$.[16] However, the buffer conditions for the two experiments are not identical, and pH has a large affect on deamidation rate. The rate of deamidation under these buffer conditions was confirmed using a synthetic peptide of the same sequence; for this peptide, the rate of deamidation was 0.0023 $min^{-1}$. The higher rate of deamidation of the synthetic peptide might reflect an association between the partially digested protein and the N-terminal peptide which introduced a degree of conformational 'freezing' of the peptide, diminishing the deamidation rate, but this remains conjecture at present.

To investigate the kinetics of both deamidation and proteolysis, a synthetic peptide of sequence VKVGVNGFGR, mass 1041.59 Da, was synthesized and was labeled at the arginine residue with both [$^{13}C_6$] and [$^{15}N_4$] giving a 10 Da mass offset relative to the natural peptide. This peptide, identical to the N-terminal peptide of GAPDH, was used to monitor the behavior of the peptide, and for quantification.[17] Because the N-terminal peptide itself contains an internal tryptic cleavage site (**VK - VGVDGFGR**), the peptide **VKVGVDGFGR** (summed across acid or amide forms) decreased slowly as digestion continued. We created a model (Figure 4) that took into account the sequential first-order processes of proteolysis ($k_1$) of the native protein ($N_{native}$) to release the amide form of the peptide (**VKVGVNGFGR**) followed by deamidation ($k_2$) to generate the acid form (**VKVGVDGFGR**).

$$N_{native} \xrightarrow{k_1} \text{VKVGVNGFGR} \xrightarrow{k_2} \text{VKVGVDGFGR}$$

Furthermore, the model also included a secondary process of proteolysis of the released peptide in either the acid or amide form to release the ValLys dipeptide. The rate of appearance of the deamidated peptide is given by

$$\text{VKVGV(N/D)GFGR} \xrightarrow{k_3} \text{VK} + \text{VGV(N/D)GFGR}$$

We assumed that the rate of deamidation ($k_2$) was independent of the N-terminal ValLys dipeptide and that the rate of tryptic removal of the N-terminal dipeptide ($k_1$) was the same, irrespective of whether the peptide was in acid or amide form. The change in amount (relative to the initial amount of protein, $N_{native(t=0)}$) of the larger peptides (**VKVGVNGFGR** + **VKVGVDGFGR**, $N + D$) as a function of time, is given by

$$N + D = N_{native}\left(\frac{k_1}{k_3 - k_1}(e^{-k_1 t} + e^{-k_3 t})\right) \quad (1)$$

As part of the same process, the shortened peptide (**VGVNGFGR** + **VGVDGFGR**, $N' + D'$) appears according to

$$N' + D' = N_{native}\left(1 - \frac{k_3}{k_3 - k_1}e^{-k_1 t} + \frac{k_3}{k_3 - k_1}e^{-k_3 t}\right) \quad (2)$$

Assuming that the rate of tryptic cleavage is consistent for both acid and amide variants, from these equations, we were able to calculate the second-order rate constants (first-order rate constant divided by protease concentration) for initial release of the large peptide ($k_1$) and the rate of proteolysis of this large peptide ($k_3$) (Figure 5). The value of $k_1$ was estimated to be $1.22 \pm 0.025$ $min^{-1} \cdot \mu M$ and for $k_3$, $0.50 \pm 0.008$ $min^{-1} \cdot \mu M$ (trypsin = 0.2 $\mu$M). As expected, the endoproteolytic release of the longer peptide is faster than the release of the N-terminal dipeptide, as trypsin is known to act poorly as a dipeptidyl peptidase. However, the release of the longer peptide is likely to be suppressed by the three-dimensional structure of the protein.

To investigate the effects of the higher order structure of GAPDH on proteolysis and subsequent deamidation, we analyzed the X-ray crystal structure of rabbit GAPDH (PDB code 1J0X.PDB). First, we used the tool NickPred,[18] which although designed to predict sites of proteolytic attack, can generate a comprehensive analysis of the environment of every residue in a protein sequence. The N-terminal region of GAPDH is rather constrained, exhibiting low temperature factors (*B*-values) and low protrusion and accessibility (results not shown). Close inspection of the structure in the vicinity of $Asn_6$ revealed this region of the polypeptide chain was folded in an extended $\beta$ configuration, constrained by 14 hydrogen bonds in a network that might be expected to constrain main chain flexibility and therefore reduce the propensity for asparagine deamidation (Figure 6). However, once the peptide was released by proteolysis, deamidation proceeded at a higher rate than that predicted from model studies. These experiments are consistent with the following propositions; that the residue in

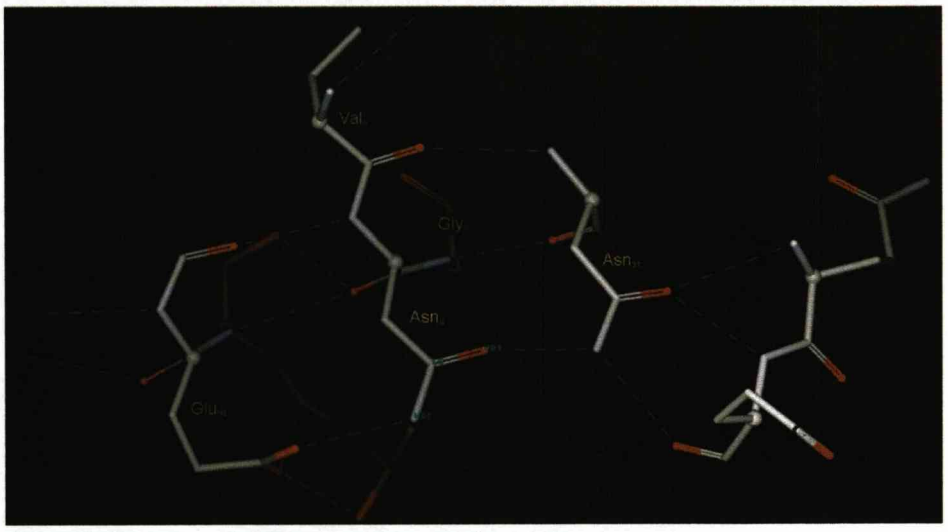



**Figure 6.** 3D structure of rabbit skeletal muscle GAPDH. X-ray crystal structure of the N-terminal region of rabbit skeletal muscle GAPDH (PDB code 1J0X) highlighting the $Asn_6Gly_7$ deamidation site and the local hydrogen bonded environment. The green dashed lines denote hydrogen bonds.

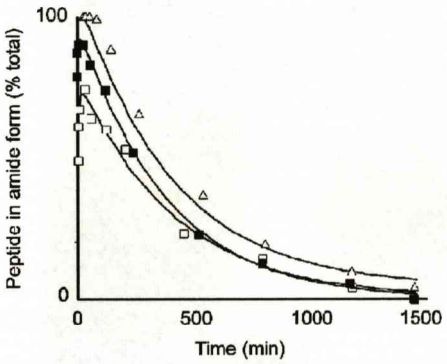



**Figure 7.** The effect of denaturing protein structure by heating on the rate of deamidation. GAPDH (1 mg/mL diluted to 0.2 mg/mL with 50 mM ammonium bicarbonate) purified from rabbit skeletal muscle (Sigma, Dorset, U.K.) was digested with trypsin in solution at a ratio trypsin/protein 1:100 at 37 °C for 24 h. Prior to digestion, GAPDH was incubated for 1 h at 4 °C (open triangles), 1 h at 60 °C (closed squares), and 1 h at 60 °C followed by 24 h at 37 °C (open squares). For each, deamidation was monitored over 24 h proteolysis and the proportion of acid and amide was calculated from the relative peak intensities of the two ions in MALDI-ToF mass spectra.

the intact protein is exclusively in the amide form, that the tryptic fragment containing the amide residue can undergo deamidation, and that deamidation is not an artifact of the mass spectrometric analysis. Excision of the peptide from the GAPDH structure relieves the constraint in the peptide backbone trajectory, permitting the deamidation reaction to take place. It followed therefore that prior denaturation of the protein might permit deamidation prior to digestion with trypsin. We conducted experiments in which we denatured GAPDH by heating to 60 °C for 1 h before proteolysis (Figure 7), a denaturation treatment that was not sufficient to cause the protein to precipitate. Subsequently, when trypsin was added, the N-terminal peptide was again released rapidly, and the proportion of amide and acid variants of the peptide was assessed as previously described. Under these circumstances, the peptide first released was approximately 80% amide, with a significant proportion of acid form being measurable. This contrasted markedly with proteolysis of the native protein, when the peptide is initially all in the amide form. We attribute this behavior to the increased conformational flexibility of the peptide in the heat-treated protein, such that the peptide could acquire a conformation that allowed deamidation. Further, this unfolded and flexible component might be expected to be hypersensitive to proteolysis and to be released first. As the digestion proceeded, additional peptide in the amide form was released, and the proportion of amide therefore increased transiently, until the deamidation reaction dominated the peptide profile. When the functions derived previously were used, we obtained a value for deamidation of 0.0023 $min^{-1}$, in close agreement with that observed previously. If the heat-treated peptide was allowed to incubate at 37 °C for 24 h after the 60 min denaturation period at 60 °C, and then proteolyzed with trypsin, the peptide first released was now only 50% in the amide form, consistent with extensive deamidation prior to proteolysis, consequential to denaturation. Again, as expected, proteolysis led to the slower release of peptide that was constrained and unable to deamidate, and there was a transient increase in the proportion of amide which again decayed at the same rate as observed previously ($k_2 = 0.0024$ $min^{-1}$). The behavior of the system was consistent with the GAPDH preparation being 76% in the amide form, and 26% in a denatured form that was then rapidly proteolyzed to generate the free acid form of the peptide. The effect of denaturation on the availability of the N-terminal peptide of GAPDH for deamidation is quite striking and defines the importance of monitoring the two processes of proteolysis and deamidation simultaneously, especially as this effect is only observed upon loss of higher order structure, and not upon increasing concentration of protease (results not shown).

## Conclusions

Deamidation is recognized as a potential source of microheterogeneity in protein structure, and it may play a significant role as a biological 'timer' that is mediated nonenzymatically.[1–5] Although many studies have emphasized the deamidation of short, flexible peptides, protein deamidation can be limited by higher order structure and might only occur at the peptide level

following proteolytic release.[8] The ease with which some peptides deamidated could then lead to the erroneous interpretation of a deamidation event as occurring in the intact protein. Difficulties of measuring deamidation have been discussed,[19] and analyses often use electrospray ionization mass spectrometry[6] and reversed-phase chromatographic matrices[8] to resolve acid and amide variants of a peptide, precluding analysis of complex mixtures. There is also considerable scope for MALDI sample ionization, which, when coupled with a simple esterification reaction, can clearly identify and characterize such deamidation variants. We suggest that there may be merit in closer examination of the isotope distribution profile of peptide mass fingerprints, to search for anomalies such as that noticed here. In particular, it is advantageous to monitor deamidation and proteolysis simultaneously when characterizing post-translational behavior of known proteins and peptides. This will also unravel information about the higher order structure of a protein, the influence of which not only on proteolysis but also on subsequent modifications to newly accessible regions of the protein, is paramount.

**Acknowledgment.** We are grateful to Dr. Gary Evans, Genus plc, for his interest in this work. This work has been supported by the BBSRC, EPSRC (EP/D013623) and Genus plc. We are grateful to Dr. D. H. Robertson for instrumentation support.

## References

(1) Robinson, A. B.; Rudd, C. J. Deamidation of glutaminyl and asparaginyl residues in peptides and proteins. *Curr. Top. Cell. Regul.* **1974**, *8* (0), 247–295.

(2) Geiger, T.; Clarke, S. Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. *J. Biol. Chem.* **1987**, *262* (2), 785–794.

(3) Friedman, A. R.; Ichhpurani, A. K.; Brown, D. M.; Hillman, R. M.; Krabill, L. F.; Martin, R. A.; Zurcher-Neely, H. A.; Guido, D. M. Degradation of growth hormone releasing factor analogs in neutral aqueous solution is related to deamidation of asparagine residues. Replacement of asparagine residues by serine stabilizes. *Int. J. Pept. Protein Res.* **1991**, *37* (1), 14–20.

(4) Deverman, B. E.; Cook, B. L.; Manson, S. R.; Niederhoff, R. A.; Langer, E. M.; Rosova, I.; Kulans, L. A.; Fu, X.; Weinberg, J. S.; Heinecke, J. W.; Roth, K. A.; Weintraub, S. J. Bcl-xL deamidation is a critical switch in the regulation of the response to DNA damage. [erratum: *Cell* **2003**, *115* (4), 503] *Cell* **2002**, *111* (1), 51–62.

(5) Weintraub, S. J.; Manson, S. R. Asparagine deamidation: a regulatory hourglass. *Mech. Ageing Dev.* **2004**, *125* (4), 255–257.

(6) Robinson, N. E.; Robinson, A. B.; Merrifield, R. B. Mass spectrometric evaluation of synthetic peptides as primary structure models for peptide and protein deamidation. *J. Pept. Res.* **2001**, *57* (6), 483–493.

(7) Doherty, M. K.; McLean, L.; Hayter, J. R.; Pratt, J. M.; Robertson, D. H.; El-Shafei, A.; Gaskell, S. J.; Beynon, R. J. The proteome of chicken skeletal muscle: changes in soluble protein expression during growth in a layer strain. *Proteomics* **2004**, *4* (7), 2082–2093.

(8) Krokhin, O. V.; Antonovici, M.; Ens, W.; Wilkins, J. A.; Standing, K. G. Deamidation of -Asn-Gly- sequeneces during sample preparation for proteomics: consequences for MALDI and HPLC-MALDI analysis. *Anal. Chem.* **2006**, *78*, 6645–6650.

(9) Stroop, S. D. A modified peptide mapping strategy for quantifying site-specific deamidation by electrospray time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 830–836.

(10) Jedrzejewski, P. T.; Girod, A.; Tholey, A.; Konig, N.; Thullner, S.; Kinzel, V.; Bossemeyer, D. A conserved deamidation site at Asn 2 in the catalytic subunit of mammalian cAMP-dependent protein kinase detected by capillary LC-MS and tandem mass spectrometry. *Protein Sci.* **1998**, *7* (2), 457–469.

(11) Nilsson, M. R.; Driscoll, M.; Raleigh, D. P. Low levels of asparagine deamidation can have a dramatic effect on aggregation of amyloidogenic peptides: implications for the study of amyloid formation. *Protein Sci.* **2002**, *11* (2), 342–349.

(12) Cournoyer, J. J.; Lin, C.; O'Connor, P. B. Detecting deamidation products in proteins by electron capture dissociation. *Anal. Chem.* **2006**, *78*, 1264–1271.

(13) Bogdan, I.; Coca, D.; Rivers, J.; Beynon, R. J. Hardware acceleration of processing of mass spectrometic data for proteomics. *Bioinformatics* **2007**, *23* (6), 724–731.

(14) Pratt, J. M.; Simpson, D. M.; Doherty, M. K.; Rivers, J.; Gaskell, S. J.; Beynon, R. J. Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protoc.* **2006**, *1* (2), 1029–1043.

(15) Rivers, J.; Simpson, D. M.; Robertson, D. H. L.; Gaskell, S. J.; Beynon, R. J. Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT. *Mol. Cell. Proteomics* **2007**, *6*, 1416–1427.

(16) Robinson, N. E.; Robinson, Z. W.; Robinson, B. R.; Robinson, A. L.; Robinson, J. A.; Robinson, M. L.; Robinson, A. B. Structure-dependent nonenzymatic deamidation of glutaminyl and asparaginyl pentapeptides. *J. Pept. Res.* **2004**, *63* (5), 426–436.

(17) Kirkpatrick, D. S.; Gerber, S. A.; Gygi, S. P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods (Duluth)* **2005**, *35* (3), 265–273.

(18) Hubbard, S. J. The structural aspects of limited proteolysis of native proteins. *Biochim. Biophys. Acta* **1998**, *1382* (2), 191–206.

(19) Lindner, H.; Helliger, W. Age-dependent deamidation of asparagine residues in proteins. *Exp. Gerontol.* **2001**, *36* (9), 1551–1563.

PR070425L

# Absolute Multiplexed Quantitative Analysis of Protein Expression during Muscle Development Using QconCAT*

Jenny Rivers‡, Deborah M. Simpson‡, Duncan H. L. Robertson‡, Simon J. Gaskell§, and Robert J. Beynon‡¶

**Stable isotope-labeled proteotypic peptides are used as surrogate standards for absolute quantification of proteins in proteomics. However, a stable isotope-labeled peptide has to be synthesized, at relatively high cost, for each protein to be quantified. To multiplex protein quantification, we developed a method in which gene design *de novo* is used to create and express artificial proteins (QconCATs) comprising a concatenation of proteotypic peptides. This permits absolute quantification of multiple proteins in a single experiment. This complete study was constructed to define the nature, sources of error, and statistical behavior of a QconCAT analysis. The QconCAT protein was designed to contain one tryptic peptide from 20 proteins present in the soluble fraction of chicken skeletal muscle. Optimized DNA sequences encoding these peptides were concatenated and inserted into a vector for high level expression in *Escherichia coli*. The protein was expressed in a minimal medium containing amino acids selectively labeled with stable isotopes, creating an equimolar series of uniformly labeled proteotypic peptides. The labeled QconCAT protein, purified by affinity chromatography and quantified, was added to a homogenized muscle preparation in a known amount prior to proteolytic digestion with trypsin. As anticipated, the QconCAT was completely digested at a rate far higher than the analyte proteins, confirming the applicability of such artificial proteins for multiplexed quantification. The nature of the technical variance was assessed and compared with the biological variance in a complete study. Alternative ionization and mass spectrometric approaches were investigated, particularly LC-ESI-TOF MS and MALDI-TOF MS, for analysis of proteins and tryptic peptides. QconCATs offer a new and efficient approach to precise and simultaneous absolute quantification of multiple proteins, subproteomes, or even entire proteomes.** ***Molecular & Cellular Proteomics 6:1416–1427, 2007.***

From the ‡Proteomics and Functional Genomics Group, Faculty of Veterinary Science, University of Liverpool, Liverpool L69 7ZJ, United Kingdom and §Michael Barber Centre for Mass Spectrometry, School of Chemistry and Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, United Kingdom

Received, December 4, 2006, and in revised form, May 16, 2007
Published, MCP Papers in Press, May 17, 2007, DOI 10.1074/mcp.M600456-MCP200

As the field of proteomics matures as a discipline, there is an increasing realization of the importance of *absolute* as well as *relative* quantification, and considerable effort is being directed toward experimental strategies to achieve this goal. Most commonly, relative protein quantification by mass spectrometry has been based on differential stable isotope labeling implemented by metabolic incorporation (1, 2) or through derivatization strategies such as ICAT (3). The mass-coded abundance tagging method (4) avoids the use of stable isotopes but requires assumptions concerning mass spectrometric response factors. To achieve relative quantification of proteins without isotope labeling or chemical modification steps, quantitative comparisons have been made of equivalent sets of mass spectrometric data by considering peptide detectability in repetitively acquired spectra or by comparing integrated extracted ion chromatograms following liquid chromatography-mass spectrometry analyses (5).

In principle, any of the approaches adopted for relative quantification may also be used for absolute quantification if reference standards are available for all analytes in known amounts. When unknowns and reference standards are co-analyzed, such approaches exploit the well established precept in analytical chemistry of internal standardization in which a known amount of a stable isotope-labeled (or otherwise differentiated) standard is added to the analyte such that the response ratio between analyte and the heavier internal standard can then be used to quantify the unknown. However, for quantification of individual proteins in a proteomics study, the true internal standard would be the corresponding protein expressed in pure and stable isotope-labeled form and quantified. This would be challenging on many fronts, including the expression of a native protein in a heterologous system to effect labeling, purification of the protein, and subsequent mass spectrometric analysis of the complex isotopic profile of the analyte and standard protein. Rather than adopt a protein-based approach, absolute quantification using proteotypic peptides as surrogates for the protein of interest has emerged using stable isotope-labeled peptide internal standards as "signature" or "proteotypic" peptides that are chemically synthesized and incorporate stable isotopes (6–9). Each protein to be quantified requires at least one stable isotope-labeled peptide that must be independently synthesized at relatively

high cost. Moreover each peptide must be separately purified and quantified (10). There is clearly a need for approaches that make this process more streamlined especially if multiple proteins are to be quantified.

We have recently introduced an efficient alternative to the chemical synthesis of multiple stable isotope-labeled peptides (11). In brief, artificial genes are designed *de novo* to direct the synthesis of novel proteins that are assemblies of signature Qpeptides derived from a number of discrete proteins. Usually these Qpeptides are arginine or lysine terminated at the C terminus as they represent and will be internal standards for tryptic peptides derived from digestion of the analyte proteins. Appropriately flanked with added features including an initiator codon, a purification tag, and protective sacrificial regions, the gene is transformed into and expressed in a heterologous system, usually bacterial. The expression strain is grown in a chemically defined medium, uniformly isotopically labeled (for example, using $^{15}NH_4Cl$ as the sole nitrogen source) or containing specific stable isotope-labeled amino acids at a high isotope enrichment such that the artificial protein becomes fully labeled. The artificial protein (termed a "QconCAT" for "quantification concatamers") is purified by virtue of the affinity tag and quantified using a suitable procedure (12). The QconCAT is mixed with a complex mixture of analyte proteins, and subsequent proteolysis releases both the stable isotope-labeled standard and the cognate peptide from the analyte. The known quantity of standard added can then be used for absolute quantification of the analyte. Because quantification of the QconCAT protein will define in absolute terms the quantity of each of the surrogate peptides, the QconCAT strategy provides an efficient means to multiplex absolute quantification. Tryptic peptides are typically 10–15 amino acids long; thus proteotypic Qpeptides from 50 proteins could be encoded in a protein comprising 500–750 amino acids. The Qpeptides are present, by design, in stoichiometrically known amounts (usually equimolar) so that each analyte peptide (and therefore protein) is simultaneously quantified.

Qpeptides are concatenated in the QconCAT protein out of their normal primary sequence context, and it is formally correct to point out that this different context could influence the quantification data (13). However, this can only occur if either the QconCAT or the analyte proteins are incompletely digested such that the yield of each peptide is incomplete. It is generally accepted that for most general proteases, such as trypsin, the $K_m$ for proteins and peptides is relatively high, and proteolytic reactions operating at substrate concentrations below this value exhibit pseudo-first order kinetics (14–16). Thus, if the rate of digestion of either the QconCAT or analyte was so low that six or seven reaction half-times could not elapse during the proteolytic reaction, there might be discordance between the yield of the standard and analyte peptide. However, the main determinant of the rate of proteolysis of native proteins is higher order structure, not primary sequence context. Tightly folded proteins, particularly those with a high proportion of $\beta$ sheet, are intrinsically resistant to proteolysis (17, 18). There is no reason, *a priori*, to expect that QconCATs would adopt such tightly folded structures. Indeed their propensity to form insoluble inclusion bodies and their recovery by dissolution in strong chaotropes both mitigate against structural impediments to proteolysis. By contrast, unless care is taken in the prior denaturation of analyte proteins, their higher order structure would almost certainly influence proteolysis and could impact absolute quantification. We stress, however, that the incomplete analyte digestion is as much an issue for quantification using synthetic peptides as those using QconCATs. We address the issue of QconCAT and analyte proteolysis here and show that it is a factor that is readily controlled.

Deployment of a QconCAT experiment has many aspects that must be optimized. We demonstrate the use of a QconCAT for absolute quantification of a group of proteins that demonstrate dramatic changes in expression during development of skeletal muscle in the chicken posthatching. We assessed the scope of the method and the magnitude and sources of variance that the method contains. We confirmed the value of guanidination (19) as a strategy to enhance peptide ion yields in MALDI-TOF MS and showed that effective quantification is attainable and equivalent in both MALDI-TOF and ESI-TOF analyses.

## EXPERIMENTAL PROCEDURES

*Materials and Reagents*—Trypsin (sequence grade) was obtained from Roche Diagnostics. All other chemicals and solvents (HPLC grade) were purchased from Sigma-Aldrich and VWR International Laboratory Supplies (Leicestershire, UK).

*Proteomics Analysis of Chicken Skeletal Muscle Soluble Fraction*—Chickens (Institut de Sélection Animale (ISA) Brown layer and Ross 308 broiler) were grown to 30 days posthatch, and animals were culled at 1, 3, 5, 10, 20, and 30 days at which time pectoralis muscle was collected (the above procedures were performed at the Roslin Institute, Edinburgh, UK). To isolate the soluble fraction of chicken skeletal muscle, 100 mg of breast tissue was homogenized in 0.9 ml of 20 mM sodium phosphate buffer, pH 7.0, containing protease inhibitors (Complete protease inhibitors, Roche Applied Science). The homogenized sample was centrifuged at 15,000 × *g* for 45 min at 4 °C. The supernatant fraction, containing soluble protein, was then removed. This was repeated, homogenizing the insoluble fraction in the same volume of sodium phosphate, and the pooled supernatant fractions were used for all analyses. The total protein concentration of each preparation was measured using a Coomassie Plus Protein Assay (Pierce).

For 1D[1] SDS-PAGE analysis, 10 $\mu$g of soluble protein samples (volume, 5–10 $\mu$l) from birds of different strains and ages were each mixed with an equal volume of reducing sample buffer (1 ml of 0.5 M Tris buffer, pH 6.8, 1 ml of glycerol, 0.02 g of SDS, 0.01g of bromphenol blue, 0.154 g of DTT) and resolved by 12.5% (w/v) SDS-PAGE prior to staining with Coomassie Blue (Bio-Safe, Bio-Rad). Gels were

[1] The abbreviations used are: 1D, one-dimensional; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; AK, adenylate kinase; CV, coefficient of variance.

destained with 10% (v/v) acetic acid, 10% (v/v) methanol.

*Preparation and Purification of QconCAT*—The artificial QconCAT gene (11) was expressed in *Escherichia coli* with a full complement of unlabeled amino acids or in the presence of [$^{13}C_6$]lysine (100 mg/liter) and [$^{13}C_6$]arginine (100 mg/liter) as the sole source of these two amino acids. Expression was induced with isopropyl β-D-thiogalactopyranoside, and the cells were harvested by centrifugation at 1400 × *g* at 4 °C for 15 min. Inclusion bodies containing QconCAT (as proven by digestion with trypsin and MALDI-TOF MS analysis; data not shown) were recovered by breaking cells using BugBuster Protein Extraction Reagent (Novagen, Nottingham, UK). Inclusion bodies were resuspended in 20 mM phosphate buffer, 6 M guanidinium chloride, 0.5 M NaCl, 20 mM imidazole, pH 7.4. From this solution, [$^{13}C_6$]lysine/arginine-labeled and unlabeled QconCAT proteins were purified separately by affinity chromatography using a nickel-based resin (HisTrap HP kit, Amersham Biosciences). Following sample loading, HisTrap columns were washed with 20 mM phosphate buffer, pH 7.4, prior to elution of the sample with the same buffer containing a higher concentration of imidazole (20 mM phosphate, 0.5 M NaCl, 500 mM imidazole, 6 M guanidinium chloride, pH 7.4) during which phase fractions (1 ml) were collected. The purified QconCAT was desalted by three rounds of dialysis against 100 volumes of 10 mM ammonium bicarbonate, pH 8.5, for 2 h using fresh buffer each time.

*Proteomics Analysis of QconCAT for Quantification of Chicken Skeletal Muscle Proteins*—The QconCAT protein was diluted to 5 μM in 50 mM ammonium bicarbonate and digested with trypsin (20:1 substrate:protease) at 37 °C for 24 h after which the digest was incubated with additional trypsin (20:1 substrate:protease) to ensure complete digestion. Peptides were analyzed by MALDI-TOF MS (M@LDI, Waters, Manchester, UK). For this, 1 μl of digested material was mixed with an equal volume of α-cyanohydroxycinnamic acid in 50% (v/v) acetonitrile, 0.1% (v/v) trifluoroacetic acid. This was allowed to dry, and peptides were acquired over the mass range 900–3000 *m/z*. For each combined spectrum, 20–30 spectra were acquired (laser energy typically 30%) with 10 shots per spectrum and a laser firing rate of 5 Hz. Data were processed using MassLynx software to subtract background noise using polynomial order 10 with 40% of the data points below this polynomial and a tolerance of 0.01. Spectral data were also smoothed by performing two mean smooth operations with a window of three channels.

*Co-digestion of QconCAT and Chicken Skeletal Muscle Soluble Proteins for Quantification*—QconCAT protein was added in a 1:10 (QconCAT:chicken skeletal muscle protein) ratio to chicken skeletal muscle soluble fraction samples taken from both broiler and layer strains at six time points during growth. For each time point, four birds were analyzed. The mixture was diluted 10-fold with 50 mM ammonium bicarbonate, and 10% (v/v) acetonitrile was added prior to addition of trypsin (20:1 substrate:protease). The reaction mixture was incubated at 37 °C for 24 h, after which the digest was incubated with additional trypsin (20:1 substrate:protease) to ensure complete digestion. 1 μl was analyzed by MALDI-TOF MS.

*Monitored Proteolysis of QconCAT and Analyte Proteins*—For QconCAT digestion, 150 μg of protein was digested with trypsin at a ratio of trypsin:protein of 1:20 and 1:100 and stopped at selected time points after addition of enzyme by removing 15 μl (containing 3 μg of protein) and adding to an equal volume of 10% (v/v) formic acid. For analyte protein digestion, 50 μg of protein was digested with trypsin at a ratio of trypsin:protein of 1:20 and stopped at 0 min, 30 min, and 24 h after addition of enzyme by removing 25 μl (containing 6 μg of protein) and adding an equal volume of 10% (v/v) formic acid. The fractions were subsequently stored at −20 °C until the end of the time course. For gel electrophoresis, fractions were dried down in a vacuum centrifuge and reconstituted in 10 μl of reducing sample buffer prior to separation by 12.5% (w/v) 1D SDS-PAGE at 200 V for 45 min. Analyte proteins were also digested in a solution containing 10% ACN (v/v) and with addition of enzyme following a 1-h incubation of the protein at 60 °C. To quantify proteolysis of analyte proteins, digestion of chicken skeletal muscle soluble proteins in solution with trypsin (as described above) was stopped at various time points during 24-h incubation at 37 °C by removing 20 μl (containing 5 μg of protein) and adding an equal volume of 10% (v/v) formic acid containing 0.5 μg of predigested QconCAT peptides. Each fraction was analyzed by MALDI-TOF MS. This experiment was repeated using protein denatured by incubation at 60 °C for 1 h prior to trypsin addition for comparison.

*Guanidination*—To enhance the signal intensity of lysine-terminated peptides in MALDI-TOF MS, lysine residues were converted to the more basic homoarginine by guanidination (20). This reaction was carried out by drying down the peptide mixture and reconstituting it in 10 μl of 7 M ammonia solution to which was added 5 μl of 0.5 M *O*-methylisourea (in double distilled $H_2O$). This was mixed thoroughly and incubated overnight at room temperature prior to drying down and desalting using $C_{18}$ ZipTips (Millipore, Watford, UK).

*LC-MS*—Peptide mixtures were analyzed by LC-ESI-Q-TOF MS using an EASY-nLC (Proxeon, Odense, Denmark) nanoflow system coupled to a Q-Tof micro (Waters). Nanoflow HPLC at 200 nl/min was used to resolve peptides (in 0.1% (v/v) formic acid) over a 50-min acetonitrile gradient (0–100%). Peptides were acquired over the mass range 400–2000 *m/z* with the capillary voltage set at 1900 V, collision energy set at 10 V, and sample cone set at 55 V for the entire 50-min gradient. The same reversed phase separation method was used to collect fractions (200 nl) directly onto a MALDI-TOF target for analysis by LC-MALDI-TOF MS.

*Assessing Analytical and Biological Variance in Quantification*—Ten identical aliquots of a chicken skeletal muscle soluble protein preparation, to each of which was added a known amount of [$^{13}C_6$]arginine/[$^{13}C_6$]lysine-labeled QconCAT, were digested in solution with trypsin and analyzed to investigate analytical variance. This was compared with biological variance (four animals at each time point) achieved through quantification by MALDI-TOF MS (with and without guanidination), LC-ESI-Q-TOF MS, and LC-MALDI-TOF MS.

*Comparison of QconCAT Method with Absolute Quantification Using a Stable Isotope-labeled Synthetic Peptide*—Quantification by the QconCAT method was also compared with that achieved using a stable isotope-labeled synthetic peptide to quantify a single analyte protein also represented in the QconCAT. The peptide of sequence LVSWYDNEFGYSNR and mass 1748.77 Da representing the abundant protein GAPDH was synthesized by Sigma-Genosys (Dorset, UK) and was labeled at the arginine residue with both $^{13}C_6$ and $^{15}N_4$ giving a 10-Da mass offset from the analyte peptide. For quantification, the synthetic peptide was added to broiler chicken skeletal muscle samples corresponding to six time points during growth with four replicate animals at each time point. Quantification data were obtained from analysis by MALDI-TOF MS using the relative intensities of the analyte and standard peaks as with QconCAT analysis.

*Investigation of the Accuracy of Quantification Using QconCAT*—Purified adenylate kinase (AK; Sigma) was added to chicken skeletal muscle soluble fraction from a 30-day broiler. AK was added from 0 to 0.02 nmol resulting in a final protein concentration of 0–300 nmol/g, and the amount of AK in the tissue was quantified by adding 0.015 nmol of QconCAT prior to digestion with trypsin. Proteolysis was allowed to continue for 24 h after which peptides were analyzed by MALDI-TOF MS.

## RESULTS

The QconCAT was designed to include surrogate peptides for 20 chicken skeletal muscle proteins. As chicken skeletal muscle matures posthatch, the protein distribution in the tis-

sue changes dramatically from a large number of proteins that are expressed in similar amounts at hatch to a relatively few high abundance proteins after 30 days of growth (Fig. 1). From previous identification studies (21), the most abundant proteins present in the soluble fraction of chicken skeletal muscle at this stage are predominantly the glycolytic enzymes. Other proteins, notably actin, have disappeared from the soluble fraction of muscle by 10 days of growth, presumably reflecting repartitioning and assembly of the myofibrillar apparatus. Finally serum proteins are detectable in muscle preparations at hatch but rapidly disappear during development. We ascribe this change to the increased exclusion of interstitial fluid as the muscle develops (22). To measure the absolute concentrations of specific proteins at various time points, we selected a group of 20 to be quantified using a single QconCAT. For each of the proteins, we chose a proteotypic peptide that gave a strong signal in previous MALDI-TOF MS analyses of tryptic digests. The peptides were used to guide construction of the DNA sequence of the QconCAT, which was synthesized, inserted into a pET21a vector, and expressed in *E. coli*. Full details of the design and expression are given elsewhere (12).

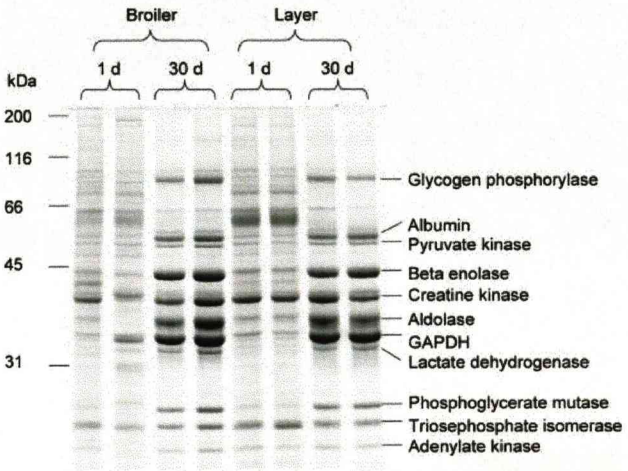



FIG. 1. **1D SDS-PAGE analysis of the soluble fraction of chicken skeletal muscle.** Two different birds were compared at 1 and 30 days (*d*) after hatch for each strain. Soluble proteins (10 $\mu$g; volume, 5–10 $\mu$l) were mixed in an equal volume with reducing SDS sample buffer, boiled for 5 min, and loaded onto a 12.5% (w/v) large format acrylamide gel prior to staining with Coomassie Blue (Bio-Safe, Bio-Rad). Gels were destained with 10% (v/v) acetic acid, 10% (v/v) methanol. Major proteins were identified by MALDI-TOF peptide mass fingerprinting.

For QconCAT expression, a typical bacterial culture of 200 ml was induced at an $A_{600}$ of 0.6–0.8, which generated 5–10 mg of QconCAT after cell breakage, recovery of inclusion bodies, and affinity chromatography of guanidinium chloride-solubilized protein on 1-ml nickel-nitrilotriacetic acid columns. After induction, the QconCAT protein was visible as a major band in 1D SDS-PAGE of a broken cell preparation (results not shown). After purification, the protein was homogeneous on 1D SDS-PAGE and was used without further purification (results not shown).

QconCAT protein was added in a 1:10 (QconCAT:chicken skeletal muscle protein) ratio to chicken skeletal muscle soluble fraction samples taken from both broiler and layer strains at six time points during growth. For each time point, four birds were analyzed. This ratio was selected pragmatically based on the abundance of the major proteins in chicken skeletal muscle soluble fraction. The influence of dynamic

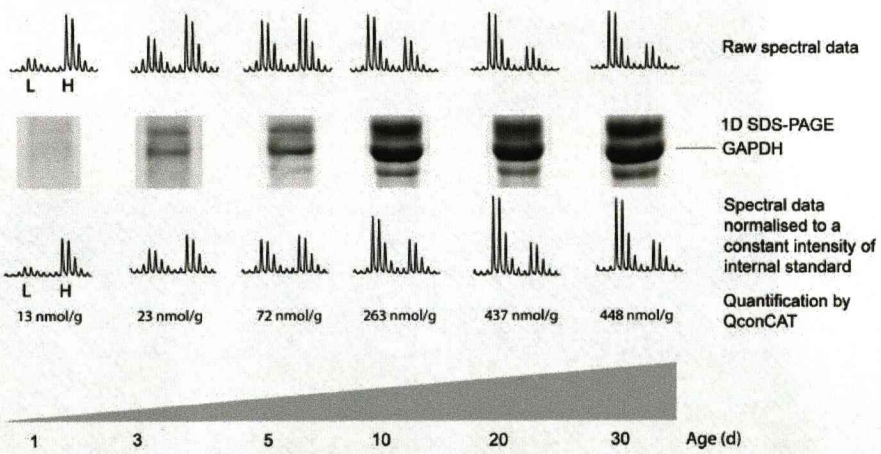



FIG. 2. **Quantification of GAPDH expression in chicken skeletal muscle.** Soluble muscle proteins were prepared from pectoralis skeletal muscle of birds from 1 to 30 days (*d*) posthatching. Each sample (70 $\mu$g of protein) was mixed with a constant amount of QconCAT (7 $\mu$g) and digested to completion with trypsin before analysis by MALDI-TOF mass spectrometry. The change in expression is measured using the relative peak intensity of the analyte and internal standard peptide at each time point. The dramatic increase in protein expression is more apparent when the spectra are normalized to a constant intensity of the internal standard. This change in protein expression is also apparent by 1D SDS-PAGE analysis of chicken skeletal muscle soluble proteins in which a constant 10 $\mu$g of protein was applied to each lane. The amount of GAPDH at each time point during growth is also expressed as nmol/g of tissue as quantified using QconCAT.

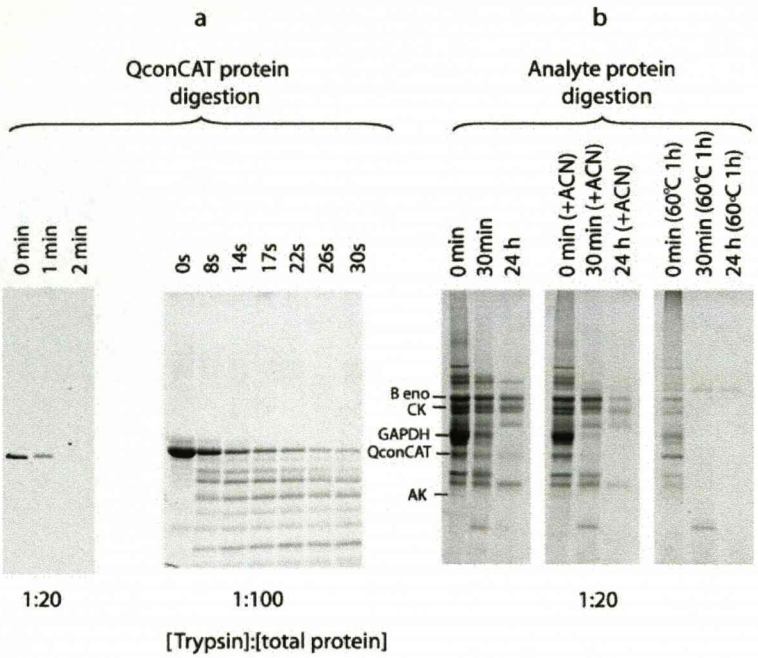



FIG. 3. **Monitored proteolysis of QconCAT and analyte proteins by 1D SDS-PAGE.** QconCAT protein (150 $\mu$g) was digested with trypsin at an enzyme: protein ratio of 1:20 and 1:00. The digestion was stopped at selected time points after addition of enzyme with 10% (v/v) formic acid. Chicken skeletal muscle soluble protein (50 $\mu$g) was digested with trypsin at an enzyme:protein ratio of 1:20 and stopped at 0 min, 30 min, and 24 h after addition of enzyme with 10% (v/v) formic acid. For gel electrophoresis, fractions from QconCAT protein digestion (*a*) and chicken skeletal muscle soluble protein digestion (*b*) were dried down in a vacuum centrifuge and reconstituted in 10 $\mu$l of reducing sample buffer prior to separation by 12.5% (w/v) 1D SDS-PAGE at 200 V for 45 min. Analyte proteins were also digested in a solution containing 10% ACN (v/v) and with addition of enzyme following a 1-h incubation of the protein at 60 °C. *CK*, creatine kinase; *B eno*, $\beta$-enolase.

range on absolute quantification of proteins in complex biological systems is discussed below. After co-digestion of chicken skeletal muscle soluble fraction and [$^{13}C_6$]arginine/lysine-labeled QconCAT, MALDI-TOF MS analysis of peptides produced highly complex mass spectra. However, 10 of 20 proteotypic peptides could be identified in the composite spectrum without further sample processing and were therefore used for quantification. For these 10 proteins, for example glyceraldehyde-3-phosphate dehydrogenase (Fig. 2), the change in protein expression can be measured during growth from 1 to 30 days posthatch by converting relative signal intensities of analyte and internal standard peptide ions into absolute quantities of analyte protein expressed as nmol/g of net weight breast muscle tissue.

The QconCAT was completely digested within 2 min such that no intermediate fragments were visible on SDS-PAGE (Fig. 3). When the trypsin was reduced to much lower levels (100:1 substrate:protease) and the digestion reaction was sampled at very short time intervals, there was some evidence for the appearance of partially fragmented intermediates, although MALDI-TOF MS analysis of these bands, once digested with trypsin, demonstrated that each "band" comprised multiple species, consistent with simultaneous tryptic attack on all scissile bonds at very similar rates. MALDI-TOF MS of peptides confirmed rapid digestion with all peptides detected within the first 2 min of digestion (data not shown).

By contrast, if the protein preparation from skeletal muscle was subjected to trypsin digestion at a ratio of 20:1 substrate: protease, many proteins were digested slowly, and even after 24 h, undigested proteins were clearly visible including $\beta$-enolase, creatine kinase, and triose-phosphate isomerase. If a low concentration (10%, v/v) of acetonitrile was included in the digestion reaction, proteolysis was faster. If the protein mixture was denatured by heating to 60 °C for 1 h before digestion, the loss of higher order structure of the substrate proteins meant that the digestion reaction was essentially complete within 30 min.

To demonstrate the importance of complete proteolysis for accurate quantification, we conducted extended digestion reactions with chicken skeletal muscle proteins from 1- and 30-day skeletal muscle. As reported previously and quantified here, these two preparations are dramatically different in the protein expression profiles (Fig. 1), providing different environments for proteolysis. The protein preparations were digested without treatment or after denaturation at 60 °C for 1 h, and the appearance of the analyte peptide used for quantification was determined by the QconCAT methodology; we have previously shown (Fig. 3) that the QconCAT was efficiently and completely digested within 2 min. In all instances, the analyte proteins were digested between 1.3 (AK) and 86 ($\beta$-enolase) times faster after denaturation, and in some instances (for example, GAPDH from 1-day muscle) the rate of digestion was very similar (Fig. 4). This is consistent with a model for proteolysis of the native protein in which the initial proteolytic attack exerts a destabilizing effect on the remaining structure such that the rate of proteolysis is increased; the initial proteolysis is effectively rate-limiting. However, in the highly specialized 30-day muscle sample, there was virtually no digestion even after 6 h of proteolysis. Indeed for all proteins studied, the rate of proteolysis of native proteins was diminished in the 30-day muscle sample; we suggest that the acute specialization of this tissue, leading to a

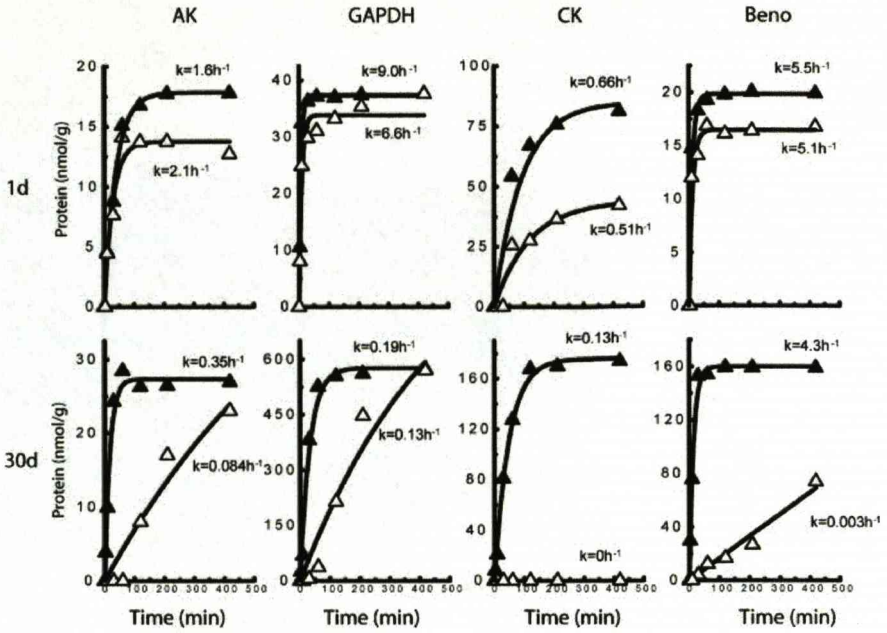
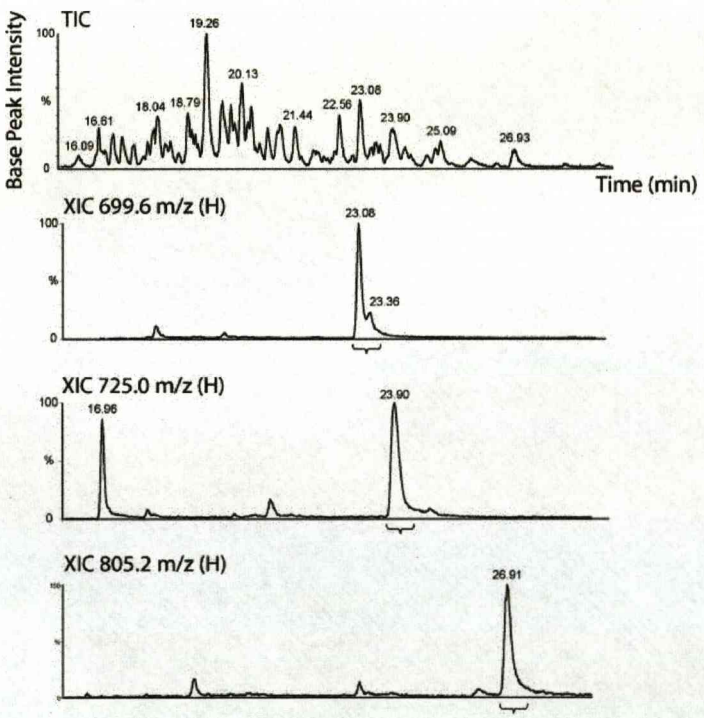



FIG. 4. **Quantification of proteolysis of analyte proteins using QconCAT.** Chicken skeletal muscle soluble protein (50 $\mu$g) was digested with trypsin at an enzyme:protein ratio of 1:20 and stopped at selected time points with 10% (v/v) formic acid and mixed with 0.5 $\mu$g of predigested QconCAT peptides for quantification. Each fraction was analyzed by MALDI-TOF MS. This experiment was repeated using protein denatured by incubating at 60 °C for 1 h prior to trypsin addition for comparison. Data are presented for four individual proteins at both 1 and 30 days (*d*) after hatch digested over 30 h (for which the first 500 min are shown) with trypsin both with (*closed triangles*) and without (*open triangles*) prior denaturation. For each, the rate constant (*k*) for digestion is expressed as $h^{-1}$. *CK*, creatine kinase; *B eno*, $\beta$-enolase.



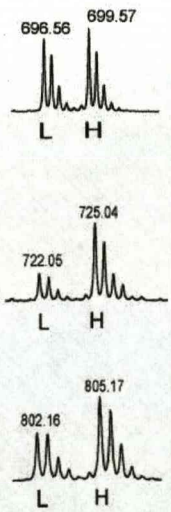




FIG. 5. **Isolation of analyte-standard peptide pairs by reversed phase chromatography.** QconCAT protein (7 $\mu$g) was added to chicken skeletal muscle soluble fraction (70 $\mu$g of protein). This mixture was digested with trypsin and analyzed by LC-ESI-Q-TOF mass spectrometry. All peptide pairs for quantification were present as doubly charged ions; there was no evidence of triply charged species. The *upper panel* is the total ion chromatogram (*TIC*; base peak intensity) of the elution profile from 16 to 29 min. The *lower panels* are the extracted ion chromatograms (*XIC*) for representative QconCAT peptides of doubly charged ions ($\beta$-enolase, 699.6 *m/z*, eluted at 23.08 min; glycogen phosphorylase, 725.0 *m/z*, eluted at 23.90 min; and triose-phosphate isomerase, 805.2 *m/z*, eluted at 25.09 min) with corresponding mass spectra showing analyte and QconCAT peptide ion pairs used for quantification presented as *insets* on the *right*. *L*, light; *H*, heavy.

predominance of relatively few proteins, might introduce other factors that impede digestion, such as aggregation into supramolecular assemblies or partial inhibition of the trypsin. In all instances, extended digestion times (greater than 24 h) resulted in complete digestion and the same quantification value irrespective of the initial state of the analyte protein preparation.

Variation in ion signal response is inherent with MALDI-TOF MS analysis (23). In particular, arginine-terminated peptides are known to yield more abundant signals than those terminated with lysine (24). In a complex MALDI-TOF mass spectrum, peptides that are abundant and have a high response factor dominate the spectrum. Theoretically proteolysis of a complex proteome (for illustration, 10,000 proteins) could

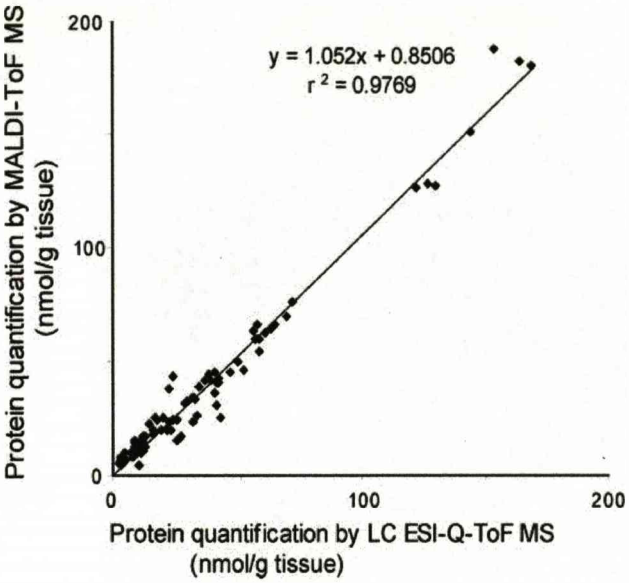



FIG. 6. **Quantification using MALDI-TOF or ESI-MS.** Soluble proteins from chicken skeletal muscle (70 $\mu$g, $n$ = 4, covering 1 to 30 days posthatch) were individually mixed with QconCAT protein (7 $\mu$g) and digested to completion with trypsin. The entire peptide mixture was analyzed by MALDI-TOF MS or by nanoflow reversed phase HPLC prior to ESI-MS, and the absolute tissue content of each of four proteins (triose-phosphate isomerase, glyceraldehyde-3-phosphate dehydrogenase, $\beta$-enolase, and $\alpha$-actin) was assessed from relative intensities of light (analyte) and heavy (standard) pairs. The absolute amount of each protein was compared using the alternative forms of mass spectrometric analysis.

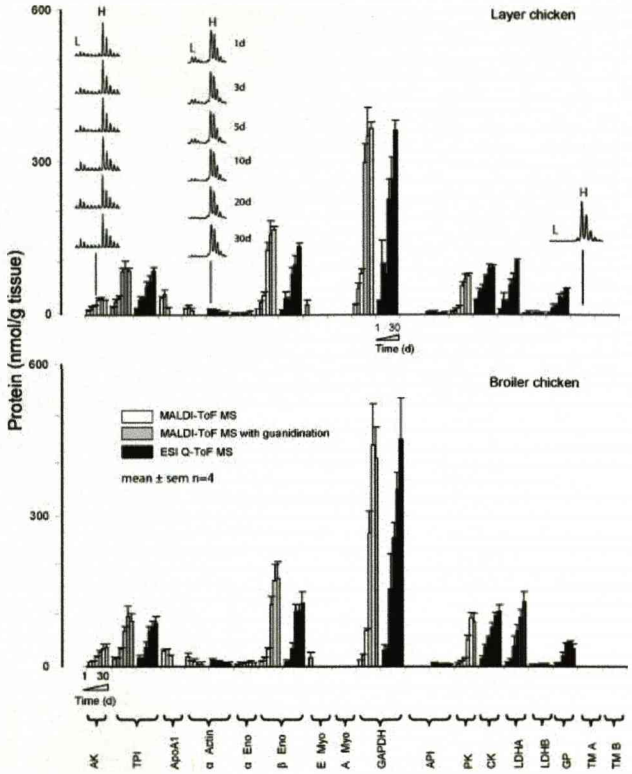



FIG. 7. **Quantification of chicken skeletal muscle protein expression by QconCAT.** Soluble protein derived from broiler and layer chickens (70 $\mu$g) was mixed with QconCAT protein (7 $\mu$g) and digested to completion with trypsin. The digests were analyzed by MALDI-TOF MS (with or without with guanidination) or LC-ESI-Q-TOF MS. For five proteins (triose-phosphate isomerase, $\alpha$-actin, $\beta$-enolase, glyceraldehyde-3-phosphate dehydrogenase, and actin polymerization inhibitor) multiple methods were used to quantify a single protein during growth; these data have been plotted separately adjacent to the alternative data set, and these have been grouped below the *x axis*. Each cluster of data represents six time points during growth (1, 3, 5, 10, 20, and 30 days (*d*)) for four birds of each strain at each time point. The data are presented as the absolute tissue amount and expressed as mean ± S.E. Mass spectra are included for proteins adenylate kinase, $\alpha$-actin, and tropomyosin A to highlight the difference in relative signal intensity. For proteins expressed as 0 nmol/g, ions corresponding to analyte peptides were not present in the spectrum (see spectral data for tropomyosin (*TM*) A). *TPI*, triose-phosphate isomerase; *Eno*, enolase; *Myo*, myosin; *API*, actin polymerization inhibitor; *PK*, pyruvate kinase; *CK*, creatine kinase; *LDH*, lactate dehydrogenase; *GP*, glycogen phosphorylase; *H*, heavy; *L*, light.

generate $10^5$–$10^6$ peptides (at ~50 tryptic peptides per protein), the dynamic range of which will be such that only the most abundant peptides and those that ionize particularly well will be identified. To achieve increased signal intensity from lysine-terminated peptides, guanidination has been used to convert lysine into the more basic homoarginine by reaction with *O*-methylisourea (20). Guanidination of a tryptic digest was effective at increasing the signal intensity of lysine-terminated peptides in the QconCAT and the analyte sample to allow quantification of two more analyte proteins by MALDI-TOF MS. To improve resolution of peptides for quantification, samples were also analyzed by LC-ESI-Q-TOF MS (Fig. 5). The alternative ionization mode coupled with the benefit of separation of peptides by reversed phase chromatography allowed quantification of a further six proteins previously not identified by MALDI-TOF MS and confirmed quantification data for many of those that had previously been analyzed. Extracted ion chromatograms for unlabeled (analyte) and labeled (QconCAT) peptides were used to locate the ions, and the chromatographic boundaries of the coincident pair of peptides were used to delineate the combined mass spectra from which peptides were quantified by mass spectrometric intensities of the doubly charged ions; there was no evidence of multiply charged ions, for example $[M + 3H]^{3+}$ corresponding to analyte-QconCAT pairs (Fig. 5). Quantification data for four proteins over 30 days of growth obtained by both methods of MALDI-TOF MS and LC-ES-Q-TOF MS showed excellent agreement such that the correlation coefficient was 0.977 (Fig. 6). All proteins that could be quantified by MALDI-TOF MS (with and without guanidination) and LC-ESI-Q-TOF MS were expressed as nmol/g of pectoralis muscle tissue. The data were obtained during growth from 1 to 30 days post-

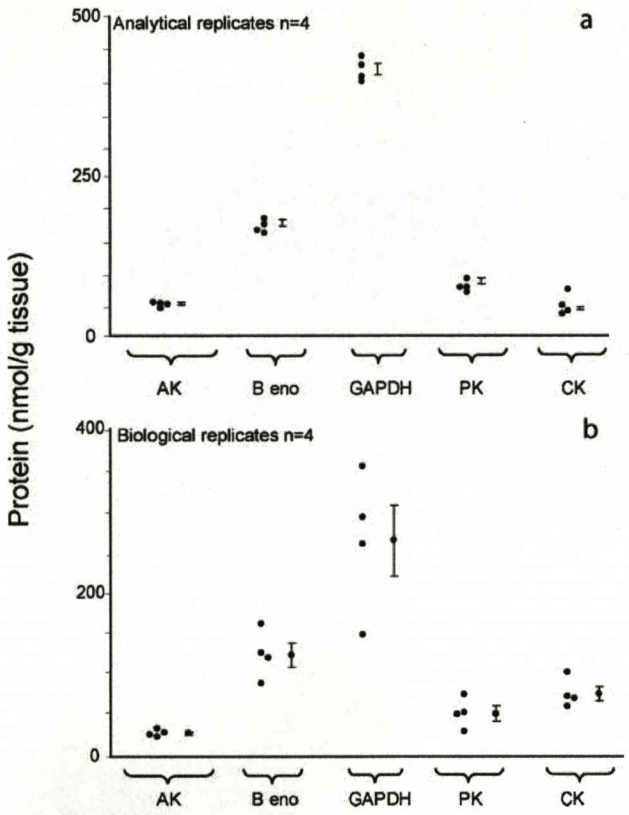



FIG. 8. **Sources of variance in a QconCAT experiment.** Soluble protein from chicken pectoralis muscle (70 $\mu$g) was mixed with QconCAT protein (7 $\mu$g) in four technically replicated experiments and digested to completion with trypsin. For each protein, individual data points are plotted to the *left* of mean ± S.E. for the same bird where $n = 4$ (*a*) and for four different birds to demonstrate biological variance (*b*). *B eno*, $\beta$-enolase; *CK*, creatine kinase; *PK*, pyruvate kinase.

hatch for four birds at each time point for chickens of the layer and broiler strains (Fig. 7). Some proteins demonstrated massive pool expansion, whereas others declined to a similar degree, covering a measurable dynamic range across all proteins of 10–550 nmol/g for a single protein (GAPDH) and as low as 2 ± 1 nmol/g ($\alpha$-enolase; 1-day broiler). Thus, in a single experiment we were able to assess protein concentrations over a 300-fold range.

To assess variance due to the analytical procedure, four identical protein mixtures (100 $\mu$l of chicken skeletal muscle (2.6 $\mu$g/$\mu$l) with 9 $\mu$l of QconCAT (2.9 $\mu$g/$\mu$l)) were digested with trypsin, and the surrogate peptides were used to quantify proteins by MALDI-TOF MS. Quantification data were collected and used to assess analytical variance (Fig. 8*a*). The reproducibility of the method was high, and the variance was similar whether four or 10 replicates were used. In both instances, the analytical variance was significantly lower than that for quantification data measured for four different birds of each strain (Fig. 8*b*). For example, the analytical variance (CV of 6.0% for $\beta$-enolase, $n = 4$) compared favorably to biological variance (CV of 24.0% for $\beta$-enolase, $n = 4$). Increasing the number of analytical replicates to 10 had very little effect on analytical variance (CV of 6.0% for $\beta$-enolase, $n = 10$; data not shown).

For some aspects of quantitative proteomics, MALDI-TOF MS has advantages. Data can be accumulated for a variable number of laser shots, ensuring comparable signal intensities between replicates. Virtually all of the signal resides in the singly charged $[M + H]^+$ ion, whereas with electrospray ionization, the signal can be distributed over a number of differently charged species. However, for complex analytical mixtures, the complexity of a MALDI-TOF mass spectrum, coupled with a noisy signal base line, can compromise quantification. One approach to simplification of a MALDI-TOF MS analysis relies on prior fractionation of the peptide mixture before deposition of successive fractions on the MALDI target (25). Chicken skeletal muscle with added QconCAT was digested and separated by reversed phase liquid chromatography, and fractions (200 nl) were collected onto a MALDI target at 1-min intervals for analysis by MALDI-TOF MS (Fig. 9). This provided an efficient detection system with peptides fixed in the solid phase for continued interrogation when acquiring data for quantification. LC-MALDI-TOF MS was used for analysis of a single chicken skeletal muscle sample to highlight the potential benefit of this method. This approach allowed quantification of the majority of proteins selected for incorporation into the QconCAT protein and consequently contributed additional information for quantification. Comparing quantification by LC-MALDI-TOF MS with both MALDI-TOF MS and LC-ESI-Q-TOF MS confirmed that all three methods of analysis give consistent and comparable quantification. This quantification can be subtle, for example in monitoring isoform changes from embryonic to adult myosin as well as a change in state from free, soluble protein to that assembled within the myofibrillar apparatus (actin). It is also possible to monitor expression of isoforms of the same enzyme for which Qpeptides differ only in a single amino acid (lactate dehydrogenases A and B).

Although there is nothing formally different between a chemically synthesized peptide and a peptide excised from a QconCAT by proteolysis, we compared the quantification of a single protein (GAPDH, which exhibits a dramatic change in abundance during posthatching development) using the QconCAT-derived peptide and the identical synthetic peptide. The correlation between data obtained using QconCAT and that obtained using the synthetic peptide was high (correlation coefficient, 0.998) (Fig. 10), and quantification data were consistent using either internal standard. A small consistent discrepancy (less than 10%) between the two methods could be attributable to the method of quantification used for the two standards. The discrepancy between the synthetic peptide and the QconCAT was reduced if we used the latter to quantify the former but was still present. We do not have an explanation for this residual discrepancy at present. We are confident, however, that the discrepancy is not attributable to

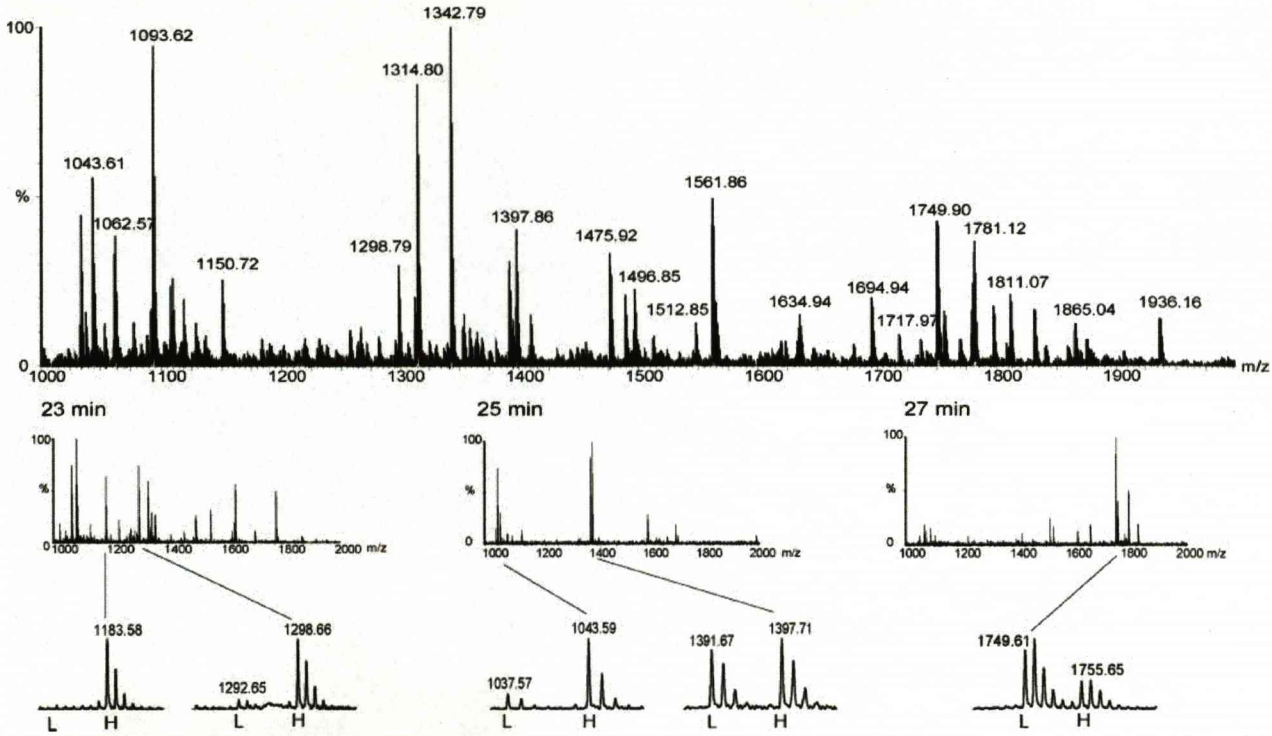



FIG. 9. **Quantification of proteins by LC-MALDI-TOF MS.** QconCAT protein (7 $\mu$g) was added to a preparation of chicken skeletal muscle soluble fraction (70 $\mu$g of protein) in a ratio of 1:10. This mixture was digested with trypsin and analyzed by LC-MALDI-TOF MS. Peptides were separated over a 50-min reversed phase acetonitrile gradient (0–100%), and fractions of 1 min (200 nl) were collected directly onto a Waters MALDI target. The *upper panel* is the MALDI-TOF mass spectrum of the entire digest; the *lower panels* illustrate three fractions collected from the reversed phase eluate at 23, 25, and 27 min. Representative pairs of analyte-standard peptides are highlighted. *H*, heavy; *L*, light.

incomplete digestion of the QconCAT (see Figs. 3 and 4). In the case of the QconCAT, we used a protein assay to determine the amount of protein as this was the same method used to quantify total protein in the analyte. For the synthetic peptide, the quantity supplied by the manufacturer is too small for independent quantification, and it was necessary to assume that the quantity in the vial was indeed that specified by the manufacturer. The difference between the two standards was minor compared with the biological variance within the system, would not contribute significant errors, and would be readily controlled by alternative QconCAT quantification strategies (see "Discussion").

To assess the accuracy of a QconCAT experiment for quantification, we spiked a known amount of AK into chicken skeletal muscle soluble fraction from a 30-day broiler. The amount of AK added was converted into protein concentration as nmol/g tissue and compared with the total concentration of AK in the tissue (nmol/g) as quantified using QconCAT (Fig. 11). As expected, there was a strong correlation ($R^2$ = 0.9992) with a slope of 1, indicating the lack of any systematic quenching effects over an extended dynamic range. Quantification of selected muscle proteins by the QconCAT strategy was also compared with densitometric quantification from 1D SDS-PAGE; the correlation of these methods was poor (data not shown; $R^2$ = 0.67), although the stain intensity was strongly proportional to the amount of protein loaded on the gel (data not shown; $R^2$ = 0.995). This is most probably due to the different affinity of individual proteins for the stain.

## DISCUSSION

QconCAT methodology has considerable potential to enhance the scope and scale of quantitative proteomics by multiplexing stable isotope dilution assays using proteotypic peptides as surrogates for the proteins of interest. The specific novelty of the QconCAT approach is derived from the efficient means of simultaneous production of multiple internal standards. Unlike chemical synthesis, biological synthesis *de novo* is not beset by "difficult" peptides (for example those with runs of serine residues or with a large hydrophobic content) that can be problematic to synthesize chemically in high purity. Moreover QconCAT proteins can be labeled using any metabolic precursor from the remarkably inexpensive uniform $^{15}N$ labeling using $^{15}NH_4Cl$ as the sole nitrogen source in the medium to specific labeling with [$^{13}C_6$]Lys/[$^{13}C_6$]Arg, which ensure that, for tryptic proteotypic peptides, each has a constant mass offset of 6 Da. Incorporation of a second labeled amino acid that is variably represented in the QconCAT can also facilitate mass isolation of the standard.

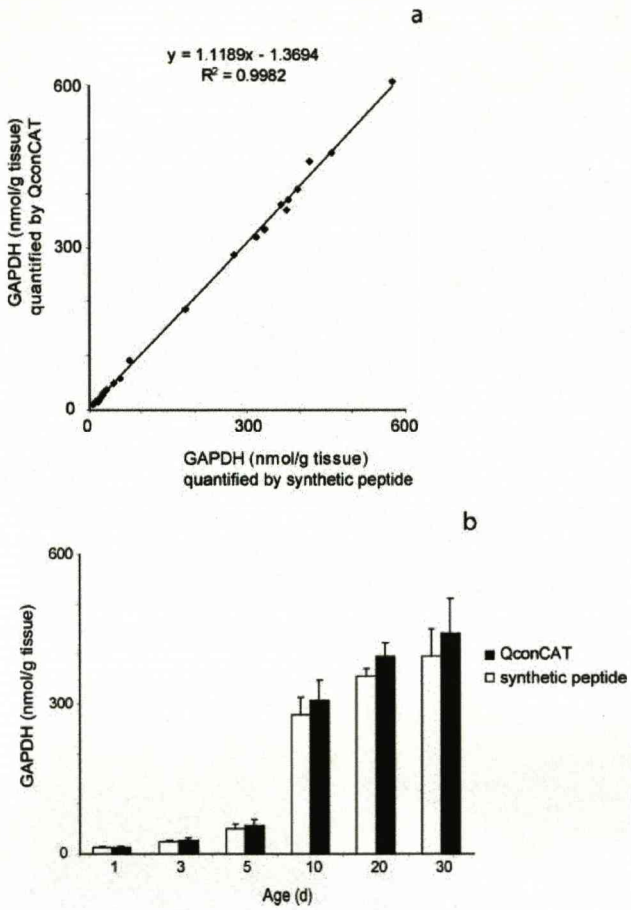



FIG. 10. **Comparison of QconCAT and synthetic peptide for quantification.** For one protein (glyceraldehyde-3-phosphate dehydrogenase), quantification was achieved relative to a QconCAT peptide and the same peptide chemically synthesized. For both methods, the internal standards (2 $\mu$g of QconCAT protein or 0.05 $\mu$g of synthetic peptide) were added to 20 $\mu$g of chicken skeletal muscle soluble protein prior to digestion with trypsin, and data were acquired using MALDI-TOF MS. The quantification by either method correlated strongly (*upper panel*). The time-dependent developmental expansion of GAPDH (nmol/g of tissue, mean ± S.E., $n$ = 4) in broiler was monitored by QconCAT or synthetic peptide (*lower panel*). *d*, days.

The imaginative application of metabolic labeling without the need for resynthesis of the QconCAT gene is an advantage of the approach that has yet to be fully exploited.

The QconCAT approach is robust to the choice of mass spectrometric method used. Each of the three methods used (MALDI-TOF MS, LC-ESI-Q-TOF MS, and LC-MALDI-TOF MS) allows quantification of individual proteins that are not detected using the alternative techniques, for example adult myosin and pyruvate kinase have only been quantified using MALDI-TOF MS, myosin-binding protein C and phosphoglycerate kinase have only been quantified using LC-MALDI-TOF MS (data not shown), and lactate dehydrogenase A has only been quantified using LC-ESI-Q-TOF MS. The ability to detect and quantify each peptide incorporated in the original QconCAT protein is very dependent on the analytical context. Although the target ionization method can influence the choice of proteotypic peptides, the opportunity remains to switch to other separation and ionization methods to gain quantification data for large numbers of proteins.

Although a chemically synthesized peptide and a QconCAT peptide are formally equivalent at the analytical stage, we compared the two approaches. Interestingly the two methods gave highly precise estimates of protein levels, but the measured values were different such that there was a consistently lower estimate of protein amount using the synthetic peptide when compared with QconCAT. Quantification of the QconCAT protein was achieved by colorimetric assay using the same method as used for the assessment of total protein concentration in the biological samples. Quantification of the synthetic peptide is based on amino acid analysis conducted by the supplier and was completed separately and prior to analysis with the mixture of analyte proteins. Indeed the quantity of the synthetic peptide supplied (five vials of 1 nmol) was sufficiently low that independent quantification by the end user could be problematical. By contrast, we routinely prepare 5–10 mg (approximately 250 nmol) of the QconCAT used here. The errors introduced by the method of standard quantification are small and, relative to the biological changes we measure here, are not significant. However, future iterations of QconCAT proteins will incorporate a common peptide for internal standard quantification by a synthetic peptide that can be labeled or unlabeled, depending on the labeling status of the QconCAT protein. This peptide, chosen because it ionizes well under MALDI or ESI, could then be used to quantify each QconCAT, normalizing all QconCAT data to a common, absolute standard. This common peptide, which is chemically synthesized, would be required in large amounts, and as such, purification and quantification of this peptide could be conducted to a very high level of confidence. By creating such a "gold standard" for quantification, data from all laboratories using QconCATs (the same or different) could be compared directly.

The application to absolute quantification of multiple proteins within complex biological systems and the adaptable nature of the QconCAT to a variety of analytical systems is clear. For development of strategies for absolute quantification, the QconCAT method provides a reproducible and relatively simple system in which multiple proteins can be quantified using alternative methods of mass spectrometry with chromatographic separation and chemical derivatization. We have made approximate estimates of the costs involved, and to quantify 50 proteins at one Qpeptide per protein, a QconCAT strategy is about 15% of the cost of comparable synthetic peptides and would yield about 250 nmol of protein compared with 5 nmol of each synthetic peptide. The error in analytical replicates is small, but there can be no "holy grail" target performance of the analytical analyses. First provided that analytical variance can be demonstrated to be substan-

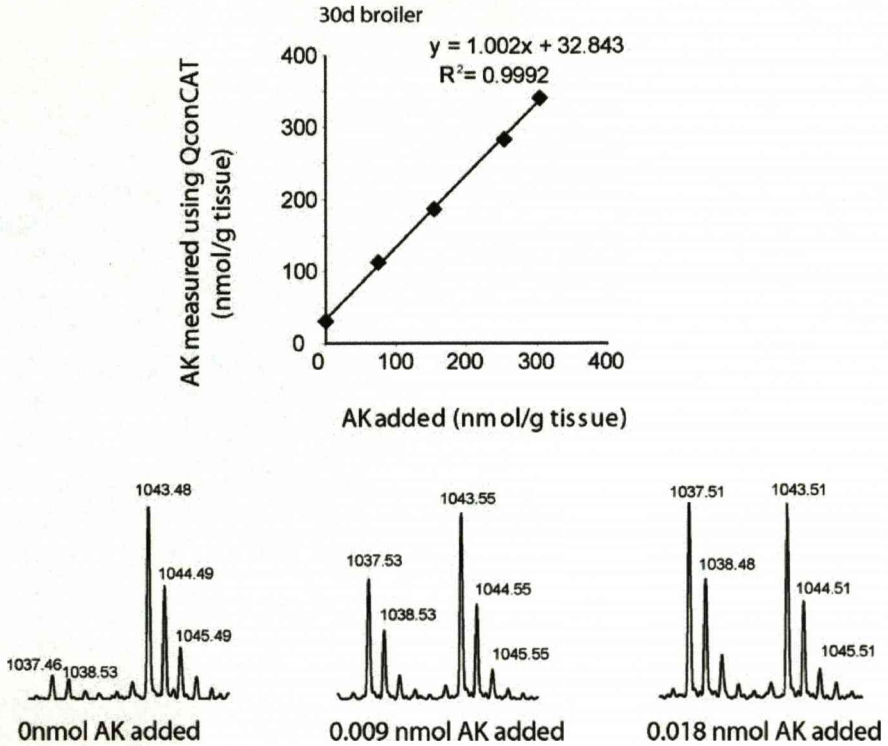



FIG. 11. **Investigation of the accuracy of quantification using QconCAT.** Purified AK was added to chicken skeletal muscle soluble fraction from a 30-day (*d*) broiler. AK was added from 0 to 0.02 nmol, which resulted in a final protein concentration of 0–300 nmol/g, and the amount of AK in the tissue was quantified by adding 0.015 nmol of QconCAT prior to digestion with trypsin. Proteolysis was allowed to continue for 24 h after which peptides were analyzed by MALDI-TOF MS. The *upper panel* shows the correlation between AK added and that quantified in the muscle using QconCAT after digestion with trypsin. Spectra showing the change in MALDI-TOF mass spectral signal intensity over the range of protein concentrations used in this experiment are shown *beneath*.

tially smaller than biological variance (as we have demonstrated here and indeed the normal expectation), it might be argued that there is a much reduced need to perform analytical replicates and that the effort should be directed toward acquisition of greater biological insight by adding new biological replicates.

For identification proteomics, little regard is paid to the completeness of the digestion of the analyte peptide; the goal is to generate sufficient peptides that are readily ionized and/or fragmented for unambiguous identification. Indeed most search engines are tolerant of and include options to match one or more than one "missed cleavage." However, when the goal shifts to the more demanding task of peptide-based quantification, it is essential that due cognizance is given to the proteolytic reactions that generate the peptides that are to be used for quantification irrespective of the method. The goal has to be complete digestion, and there are several approaches that can be taken to ensure that this has occurred. This should also be checked experimentally. It would also be feasible to embed two Qpeptides for each protein in a single QconCAT or even two different QconCATs to enhance confidence, but we do not subscribe to the view that this is necessary in many instances. Finally our experience with a large number of QconCATs[2] is that they are proteolyzed at rates that are far higher than analyte proteins. In all of these QconCAT constructs, we made no attempt to preserve the primary sequence context of the Qpeptides, and it is clear that this is not an important factor in QconCAT design; the selection of suitable proteotypic peptides in the design phase is much more critical. In this regard, the recent work by Aebersold and co-workers (26) points the way toward more effective nomination of Qpeptides for QconCATs.

*Acknowledgments*—We thank the Roslin Institute, Edinburgh, UK for growing birds and collecting muscle samples. We are grateful to Dr. Dominique Rocha, Genus plc, for interest in this work.

* This work was supported by Biotechnology and Biological Sciences Research Council Grant BB/C007433/1 (to R. J. B. and S. J. G.) and by Genus plc. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ To whom correspondence should be addressed: Proteomics and Functional Genomics Group, Faculty of Veterinary Science, University of Liverpool, Crown St., Liverpool L69 7ZJ, UK. Tel.: 44-151-794-4312; Fax: 44-151-794-4243; E-mail: r.beynon@liv.ac.uk.

[2] J. Rivers, D. M. Simpson, D. H. L. Robertson, S. J. Gaskell, and R. J. Beynon, unpublished observations.

REFERENCES

1. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
2. Sechi, S., and Oda, Y. (2003) Quantitative proteomics using mass spectrometry. *Curr. Opin. Chem. Biol.* **7,** 70–77
3. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999

4. Cagney, G., and Emili, A. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.* **20,** 163–170
5. Higgs, R. E., Knierman, M. D., Gelfanova, V., Butler, J. P., and Hale, J. E. (2005) Comprehensive label-free method for the relative quantification of proteins from biological samples. *J. Proteome Res.* **4,** 1442–1450
6. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 6940–6945
7. Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W., and Guild, B. (2004) Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and $^{13}C$-labeled peptide standards. *Proteomics* **4,** 1175–1186
8. Ishihama, Y., Sato, T., Tabata, T., Miyamoto, N., Sagane, K., Nagasu, T., and Oda, Y. (2005) Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards. *Nat. Biotechnol.* **23,** 617–621
9. Kirkpatrick, D. S., Gerber, S. A., and Gygi, S. P. (2005) The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods (Duluth)* **35,** 265–273
10. Pan, S., Zhang, H., Rush, J., Eng, J., Zhang, N., Patterson, D., Comb, M. J., and Aebersold, R. (2005) High throughput proteome screening for biomarker detection. *Mol. Cell. Proteomics* **4,** 182–190
11. Beynon, R. J., Doherty, M. K., Pratt, J. M., and Gaskell, S. J. (2005) Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* **2,** 587–589
12. Pratt, J. M., Simpson, D. M., Doherty, M. K., J., R., Gaskell, S. J., and Beynon, R. J. (2006) Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protocols* **1,** 1029–1043
13. Kito, K., Ota, K., Fujita, T., and Ito, T. (2007) A synthetic protein approach toward accurate mass spectrometric quantification of component stoichiometry of multiprotein complexes. *J. Proteome Res.* **6,** 792–800
14. Hubbard, S. J., and Beynon, R. J. (2001) Proteolysis of native proteins as a structural probe, in *Proteolytic Enzymes. A Practical Approach* (Beynon, R. J., and Bond, J. S., eds) pp. 233–264, Oxford University Press, Oxford
15. Hubbard, S. J. (1998) The structural aspects of limited proteolysis of native proteins. *Biochim. Biophys. Acta* **1382,** 191–206
16. Zappacosta, F., Pessi, A., Bianchi, E., Venturini, S., Sollazzo, M., Tramontano, A., Marino, G., and Pucci, P. (1996) Probing the tertiary structure of proteins by limited proteolysis and mass spectrometry: the case of Minibody. *Protein Sci.* **5,** 802–813
17. Hubbard, S. J., Beynon, R. J., and Thornton, J. M. (1998) Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.* **11,** 349–359
18. Wu, C., Robertson, D. H., Hubbard, S. J., Gaskell, S. J., and Beynon, R. J. (1999) Proteolysis of native proteins. Trapping of a reaction intermediate. *J. Biol. Chem.* **274,** 1108–1115
19. Brancia, F. L., Oliver, S. G., and Gaskell, S. J. (2000) Improved matrix-assisted laser desorption/ionization mass spectrometric analysis of tryptic hydrolysates of proteins following guanidination of lysine-containing peptides. *Rapid Commun. Mass Spectrom.* **14,** 2070–2073
20. Hale, J. E., Butler, J. P., Knierman, M. D., and Becker, G. W. (2000) Increased sensitivity of tryptic peptide detection by MALDI-TOF mass spectrometry is achieved by conversion of lysine to homoarginine. *Anal. Biochem.* **287,** 110–117
21. Doherty, M. K., McLean, L., Hayter, J. R., Pratt, J. M., Robertson, D. H., El-Shafei, A., Gaskell, S. J., and Beynon, R. J. (2004) The proteome of chicken skeletal muscle: changes in soluble protein expression during growth in a layer strain. *Proteomics* **4,** 2082–2093
22. McLean, L., Doherty, M. K., Deeming, D. C., and Beynon, R. J. (2004) A proteome analysis of the subcutaneous gel in avian hatchlings. *Mol. Cell. Proteomics* **3,** 250–256
23. Baumgart, S., Lindner, Y., Kuhne, R., Oberemm, A., Wenschuh, H., and Krause, E. (2004) The contributions of specific amino acid side chains to signal intensities of peptides in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **18,** 863–868
24. Krause, E., Wenschuh, H., and Jungblut, P. R. (1999) The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal. Chem.* **71,** 4160–4165
25. Mirgorodskaya, E., Braeuer, C., Fucini, P., Lehrach, H., and Gobom, J. (2005) Nanoflow liquid chromatography coupled to matrix-assisted laser desorption/ionization mass spectrometry: sample preparation, data analysis, and application to the analysis of complex peptide mixtures. *Proteomics* **5,** 399–408
26. Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold, R. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25,** 125–131

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 6 2007, pages 724–731
doi:10.1093/bioinformatics/btl656

*Gene expression*

# Hardware acceleration of processing of mass spectrometric data for proteomics

Istvan Bogdan[1], Daniel Coca[1,*], Jenny Rivers[2] and Robert J Beynon[2]

[1]Department of Automatic Control & Systems Engineering, The University of Sheffield, Mappin Street, Sheffield S1 3JD, UK, and [2]Proteomics and Functional Genomics Group, Faculty of Veterinary Science, University of Liverpool, Crown Street, Liverpool L69 7ZJ, UK

Received on September 25, 2006; revised on November 23, 2006; accepted on December 21, 2006

Advance Access publication February 3, 2007

Associate Editor: Golan Yona

**ABSTRACT**

**Motivation:** High-resolution mass spectrometers generate large data files that are complex, noisy and require extensive processing to extract the optimal data from raw spectra. This processing is readily achieved in software and is often embedded in manufacturers' instrument control and data processing environments. However, the speed of this data processing is such that it is usually performed off-line, post data acquisition. We have been exploring strategies that would allow real-time advanced processing of mass spectrometric data, making use of the reconfigurable computing paradigm, which exploits the flexibility and versatility of Field Programmable Gate Arrays (FPGAs). This approach has emerged as a powerful solution for speeding up time-critical algorithms. We describe here a reconfigurable computing solution for processing raw mass spectrometric data generated by MALDI-ToF instruments. The hardware-implemented algorithms for de-noising, baseline correction, peak identification and deisotoping, running on a Xilinx Virtex 2 FPGA at 180 MHz, generate a mass fingerprint over 100 times faster than an equivalent algorithm written in C, running on a Dual 3 GHz Xeon workstation.

**Contact:** D.Coca@sheffield.ac.uk

## 1 INTRODUCTION

The phenomenal advances in proteomics that have been made in recent years are readily attributed to advances in mass spectrometry (MS), notably soft ionization modes and tandem instrumentation, coupled with new tools for processing spectral data and database searching. The sensitivity and selectivity of the current generation of mass analysers is notable, and useable mass spectra can be recovered from vanishingly small amounts of material. Perhaps the simplest MS method in proteomics is that of peptide mass fingerprinting (PMF). PMF is a protein identification technique in which a protein is proteolyzed using an endopeptidase of defined specificity (usually trypsin) and the masses of the ensuing limit peptide fragments are measured. The proteins are identified by matching the measured molecular masses of these peptide fragments against theoretical peptide mass profiles generated from protein sequence database. PMF is readily delivered at high sensitivity through routine instrumentation such as MALDI-ToF mass spectrometers and although tandem MS approaches can recover more information from single peptides, PMF still plays an important role. Indeed, as more genomes are sequenced, and cross-species matching methods are developed, PMF may assume greater importance for many sub-proteome studies.

PMF involves two basic operations. The first is processing of the raw mass spectrum to derive a mass fingerprint, generating a data set in which the only variable is the mass of each peptide (relative intensities of different ions are not routinely used in PMF). When the mass spectrum has been processed, the list of masses are first filtered to remove spurious masses such as those derived from trypsin or matrix clusters, and the remaining peptide masses constitute the fingerprint that is used to search the protein databases for a possible match. A correlation score is computed between the database entries and the unknown peptide fragment mass list. The matches with the highest score form the final candidate protein list to be returned to the user.

At present, the time required for processing of the raw mass spectrum and the subsequent database search can exceed that of acquisition of the mass spectrometric data, especially by MALDI-ToF, which boasts acquisition rates of up to 200 spectra/s. If PMF is to remain as a key method in proteomics, one compelling gain would be a system in which the raw spectra are processed and searched against the protein database in the same time frame as acquisition—this would give 'real-time PMF' (RTPMF). However, for the goal of RTPMF to be realized, there remains the need for substantial acceleration of these two stages (spectrum processing and database searching).

A very effective approach to speed up the computations is based on the development of dedicated hardware processors that are optimized to perform specific algorithms. The acceleration, compared with the standard sequential microprocessor, is achieved by concurrent implementation of different arithmetic and logic operations that make up a computational loop and by concurrent execution of several computation loops. A major drawback of this approach used to be the prohibitive costs associated with manufacturing a dedicated integrated circuit (ASIC).

*To whom correspondence should be addressed.

However, the hardware implementation has become a much more cost effective solution due to the availability of high-density field programmable gate arrays (FPGAs) and of high-level system design and development tools, which make possible the implementation of very complex hardware designs with almost the same ease as the software implementation. An FPGA is a large-scale integrated circuit that can be programmed (and re-programmed) after it has been manufactured. Early attempts to use FPGA devices in bio-computation were made to accelerate gene sequence analysis (Fagin *et al.*, 1993). FPGAs, which are well suited for high-performance, high-bandwidth and parallel processing applications, have been successfully employed to speed up DNA sequencing algorithms (Hughey 1996; Guerdoux-Jamet *et al.*, 1997; Wozniak 1997; Lavenier, 1998; Guccione *et al.*, 2002; Simmler *et al.*, 2004). FPGAs were also used in the attempt to accelerate search of substrings similar to a template in a proteome (Marongiu *et al.*, 2003). More recently, FPGAs have been used to accelerate sequence database searches with MS/MS-derived query peptides (Anish *et al.*, 2005). This hardware-based solution can reportedly locate a query within the human genome about 32 times faster than a software implementation running on a 2.4 GHz processor. A hardware sequence alignment tool implemented in FPGA is also available (Oliver *et al.*, 2005).

In addition to developing approaches for real-time database searching, we have implemented FPGA solutions for processing of raw mass spectra. This article describes the design and hardware implementation of a mass spectrum processor which performs all computational tasks involved in the generation of a mass signature from a raw spectrum namely, smoothing, peak detection and the coalescence of natural isotopomers into a single mass ('deisotoping'). The mass spectrum processor, which is implemented on a Xilinx XC2V8000 FPGA and runs at 180 MHz, achieves more than 100-fold speed-up compared with a C software implementation running on a dual 3 GHz Xeon Server with 4 GBytes of memory.

## 2 METHODS

A MALDI-ToF mass spectrum of a typical tryptic digest of a protein generates pairs of mass-to-charge (m/z) and abundance values. Typically, the number of points in the spectrum ranges from a few thousand to a few hundred thousand. The determination of experimental peptide masses (the so called peptide mass fingerprint) requires relatively complex processing of the raw mass spectrum in order to discriminate between spectral peaks that correspond to digested peptides and the associated isotopomer peaks and the spurious peaks caused by noise and sample contamination.

To create the raw data used to evaluate the FPGA implementation, single proteins and complex protein mixtures were diluted with 50 mM ammonium bicarbonate and digested with trypsin at a ratio of protein: enzyme of 50:1. One protein that was used was an artificial QconCAT protein chosen designed so that all tryptic fragments fell within the range 1000–2500 m/z (Beynon *et al.*, 2005; Pratt *et al.*, 2006). Digestion was carried out at 37°C for 24 h after which time, 1 µl digested material was spotted onto a MALDI target. This was mixed with 1 µl α-cyano hydroxycinnamic acid matrix and analysed using a Micromass M@LDI mass spectrometer (Waters, Manchester, UK) typically over the m/z range 800–4000.

The FPGA spectra processor was designed to implement, with some variations, the algorithm proposed by Samuelsson *et al.* (2004). The major difference is the method used by the FPGA processor to implement aggregation of natural isotopomers (due primarily to the natural abundance of $^{13}C$ and $^{15}N$)—the algorithm implemented in FPGA uses Poisson distributions to approximate the isotopic patterns for every peptide (Breen *et al.* 2000).

The algorithm described in Samuelsson *et al.* (2004) has several steps. First the baseline and noise levels are estimated over an arbitrary interval adjusted by user parameters, ($\omega$ and $\Omega$) which divides the raw spectrum. The data points in the spectrum are classified compared with the level of noise and the baseline into noise, baseline and signal points. Then, peaks are constructed from a group of data points where at least one point has to be signal. In the next step the constructed peaks are grouped into clusters that are further processed to identify the monoisotopes.

The deisotoping algorithm proposed by (Breen *et al.*, 2000), was preferred for FPGA implementation. Here Poisson modelling is applied to determine monoisotopic masses (deisotoped peaks) from isotopically resolved groups (clusters). The abundances of the higher isotopic contributions for a monoisotopic peak are computed using Poisson distribution models that have been shown to match very well theoretical distributions (Breen *et al.*, 2000).

A good test of such an algorithm is in the deconvolution of the overlapping mass spectra of a peptide containing an asparagine residue ('amide') and its cognate acid product in which the side chain amide residue has been deamidated. The two mass spectra overlap by 1 Da, and effective deconvolution would be able to apportion the signal into the relative proportions of acid and amide.

The block diagram of the hardware processor is depicted in Figure 1. The implementation has two major functional blocks: a peak detection unit, which identifies all significant spectral peaks and a peptide identification unit that generates the final list of peptide masses and associated abundances.

The peak detection unit implements smoothing, baseline and noise level estimation in order to discriminate between signal and noise peaks. The first block implements a Savitzky–Golay smoothing filter (Savitzky *et al.*, 1964) with a user-defined window that can be chosen according to the instrument resolution (number of data points recorded per 1 Da). The smoothing operation is optional; the user can specify if the data is pre-processed or not. The Savitzky–Golay smoothing operation is implemented as a convolution of the filter coefficients with the abundance input stream. A delay equal to the filter latency (DELAY A in Fig. 1) is applied to the mass values data stream in order to preserve synchronicity between the mass and abundance values. The number of coefficients depends on the chosen window size. The maximum window size allowed is 43 which correspond to $43 \times 43$ coefficients. The coefficient matrix is loaded into the FPGA memory before processing operation starts. The smoothing operation is implemented as a single channel parallel filter using a Xilinx Logicore block (Xilinx, 2004). Smoothing improves the shapes of peaks which helps peak detection but slightly degrades the abundance values. Minor corrections can however be implemented to compensate the small diminution in abundance. Figure 2 shows the effect of smoothing a segment of real data with a Savitzky–Golay filter that has a polynomial order of 11 and window (or frame) size of 23.

The Y, Z, W computation block computes the minimum (Z), maximum (Y) abundances and their difference ($W = Y - Z$) over a small sliding window of maximum length of $\omega = 16$ points, as described in Samuelsson *et al.*, (2004). It has a structure similar to that of a median filter that sorts in ascending order its input data stream over its filter length. Instead of computing the median, the maximum and minimum values are found.

The baseline and noise estimation block uses Y,Z,W values to compute the baseline (Y base) and noise level (Y noise) over a larger moving window of length $\Omega \times \omega$ points, where $\Omega$ is a user-defined

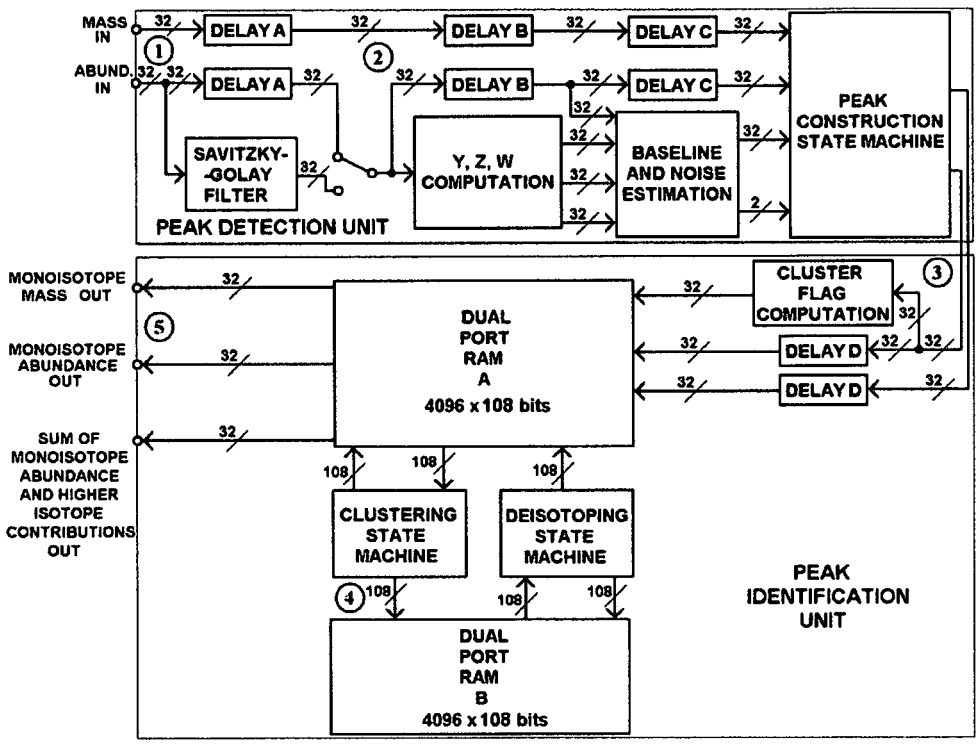



Fig. 1. Block diagram of the mass spectra processor implemented in FPGA.

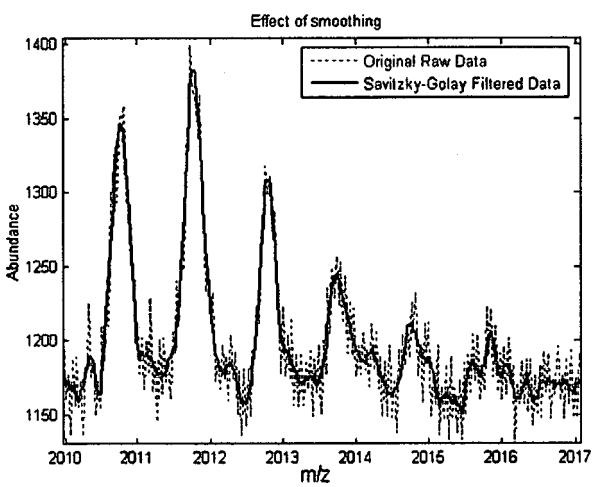



Fig. 2. Raw (dotted line) and Savitsky–Golay smoothed (solid) mass spectrum segment.

parameter (Samuelsson *et al.*, 2004). The (Y noise) and (Y base) values are used to compute the signal-to-noise ratio ($S_n$) for every data point in the spectrum. This is compared with a threshold ($\tau_1$), which can be set by the user. Each abundance value is classified as noise, support or signal depending on whether the associated $S_n$ value is $S_n < 1$, $1 < S_n < \tau_1$ or $S_n > \tau_1$, respectively. The outputs of the baseline and noise estimation block are a 2-bit classification flag (flag = 1-noise, flag = 2-support and flag = 3-signal) and the estimated baseline (32 bits) calculated for each abundance value. This data stream is aligned with the original mass-abundance pairs of the input spectrum (Fig. 1).

The peak construction state machine generates a list of valid peaks based on the 2-bit classification flag. A peak is defined as the set of mass/abundance data pairs that are either support or signal (flag = 2 or flag = 3) and are bounded by noise (flag = 1). The mass and abundance associated with each identified signal peak are calculated as the centred mass and the (baseline subtracted) peak maximum, respectively. The effect of the baseline subtraction is shown in Figure 3. The resulting peak list is written in a dual port RAM block for further analysis.

The peptide identification unit consists of a clustering and a deisotoping unit. The peaks in a cluster correspond to the isotopes of one or more singly charged chemical compounds, separated by the mass of a neutron. Clustering involves grouping together peaks so that the m/z distance between two successive peaks is between $1 - \tau_2$ and $1 + \tau_2$ where $\tau_2$ is another user selectable value, typically set to 0.2 (Samuelsson *et al.*, 2004).

The first block of the clustering unit ('Cluster flag computation') computes the distance between all consecutive signal peaks from a distance of $1 + \tau_2$ starting with the lowest mass value $m_1$. To speed up computations, there are $p$ circuits that compute mass differences $m_2 - m_1, \ldots, m_{p+1} - m_1$ between $m_1$ and the following $p$ consecutive mass values $m_1 < m_2 < \ldots < m_p < m_{p+1}$ in parallel. In our design, $p$ is an adjustable parameter, which is selected according to mass spectrometer resolution, to be equal to the maximum number of signal peaks that are registered within a window of $1 + \tau_2$. The parallel processing of these peaks is implemented by a FIFO (first in first out queue) structure of length $p$. The data flow through this circuit is depicted in Figure 4.

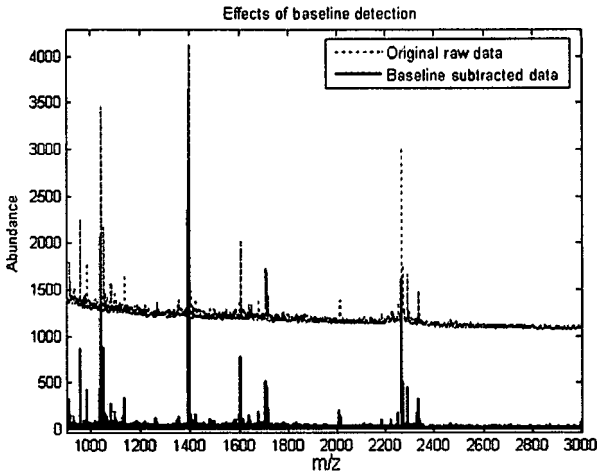



Fig. 3. Effects of baseline detection.

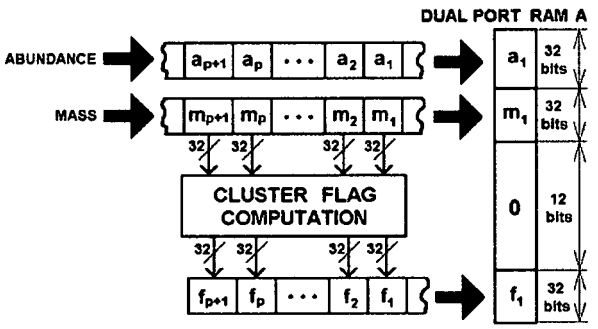



Fig. 4. Data flow for cluster flag generation.

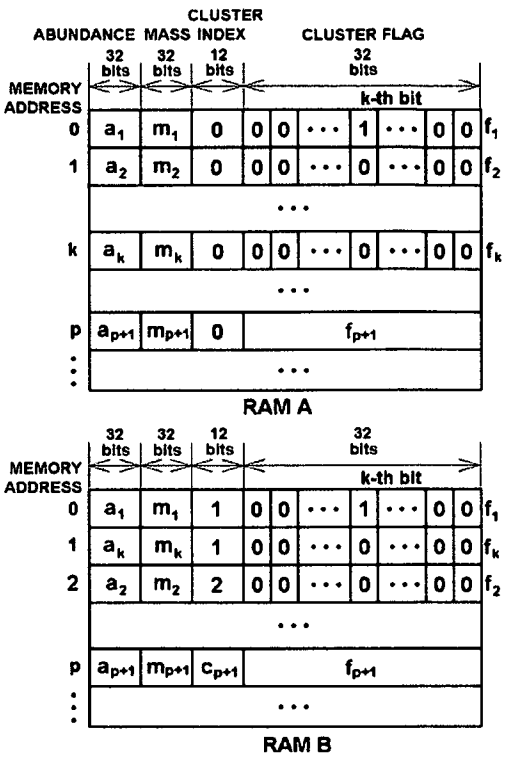



Fig. 5. Memory operations during clustering.

This is the only example of explicit algorithm parallelization. The main approach used to speed-up computations is through instruction pipelining. This technique is particularly suited for mass spectra processing where the same sequence of operations is applied to a long data stream.

In the current configuration, for each mass value that is being processed, the clustering unit generates a $p$-bit cluster flag $f$. Typically about 50–100 samples per m/z unit are taken, so the FIFO length is greater or equal to the number of possible peaks in 1.2 m/z distance. Assuming that at least three measurement points define a signal peak, and there are 100 measurements in a unit of m/z, the maximum number of constructed peaks in one m/z unit is 33 and in 1.2 m/z unit is 40. In the current design, $p$ is 32.

If the distance between $m_1$ and $m_k$ is within the range $1 \pm \tau_2$, the $k$-1 bit of this word is set to 1 indicating that $m_k$ is a potential isotopomer of $m_1$. The mass values, abundance values and associated cluster flags are concatenated and stored in RAM (A) at consecutive memory locations (increasing mass) as a 32+32+32+12 bit word (32 for mass, 32 for abundance, 32 cluster flag and 12 bits reserved for further processing) as shown in Figure 5. These records are processed to group signal peaks into clusters. If all the bits of the $f_1$ flag corresponding to $m_1$ are zero this indicates that $m_1$ has no isotopes. If the $k$th bit of this word is 1 this indicates that the peak corresponding to the $m_{k+1}$ mass value is part of the cluster having $m_1$ as the monoisotopic ion. Next, by analysing the index $f_k$ associated with $m_k$, it is possible to identify other peaks that belong to the same cluster. The process continues until the cluster flag associated with the last signal peak added to the cluster has only zero entries.

Clustering is implemented as a state machine that sequentially analyses the data stored in the first dual port RAM(A) and generates clusters that are stored in the second dual-port RAM(B). Each RAM block was configured to have 432 Kbits storage space (4 Kb address space and 108 bits word length) which can store 4096 peaks. The cluster flag associated with the latest signal peak added to the cluster is used to compute the memory address of the next peak to be included in the cluster from RAM(A). The peaks identified as belonging to one cluster are stored at consecutive memory locations in RAM(B). Each cluster is also indexed with a number of 12 bits called cluster index, a unique identification value for every cluster, stored with the member peaks. The memory content of RAM(A) before clustering and RAM(B) after clustering is shown in Figure 5. The first peak from address 0 has a single bit set to one at the $k$-th position of its cluster flag. This indicates that the peak stored at address $k$ is part of the same cluster as the first peak. As a result the peaks 1 and $k$+1 will be stored in RAM(B) at successive memory locations. In addition, to indicate that they are part of the first cluster, both peaks will have their 12 bits cluster indexes set to 1, The cluster will not include more peaks because the peak at address $k$ has its cluster flags null. The clustering process continues until all the peaks from RAM(A) are visited. The result in RAM(B) will be the same list of peaks, this time ordered by increasing cluster indexes and increasing mass values for peaks belonging to the same cluster.

For example, a cluster of eight peaks is shown in Figure 6. The peaks are situated at ~1 Da distance with m/z values: $m_1 = 2265.35$,

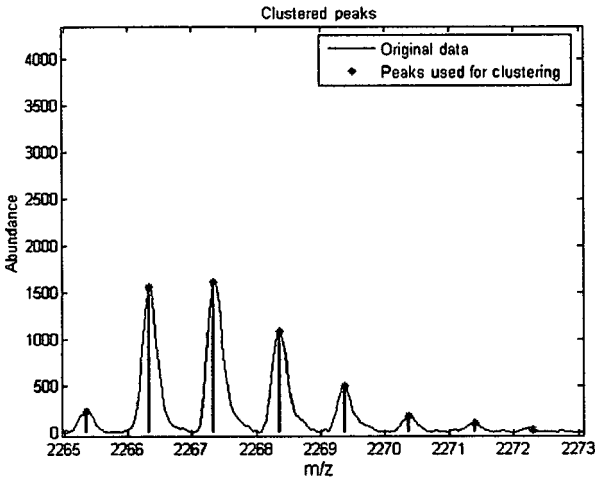



Fig. 6. Results of clustering.

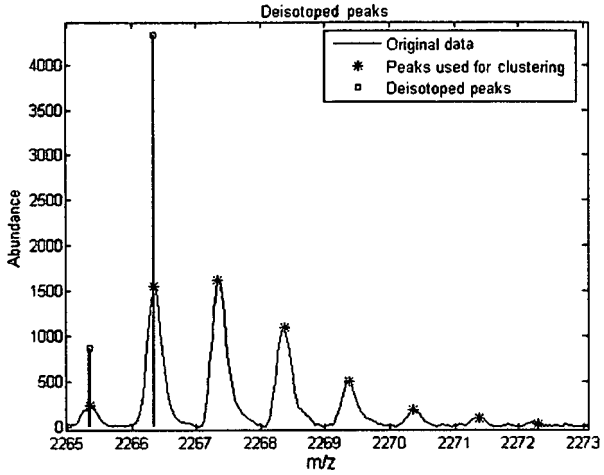



Fig. 7. Deisotoping results. The cluster of peptide ions in the Figure are deconvoluted into two species with overlapping isotopomer profiles.

$m_2 = 2266.35$ Da, $m_3 = 2267.34$ Da, $m_4 = 2268.37$ Da, $m_5 = 2269.37$ Da, $m_6 = 2270.36$ Da, $m_7 = 2271.39$ Da, $m_8 = 2272.29$ Da. In practice such a cluster has to be processed further since one cluster may contain more than one monoisotope, that is, the peaks in a cluster can be viewed as a superposition of isotopic distributions.

The deisotoping unit isolates all monoisotopic masses in a cluster and calculates the total abundance of the peptide by summation of the intensities of all of the isotopomers. The hardware algorithm implements deisotoping based on an approximation of isotopic patterns by a Poisson distribution as proposed by Breen *et al.*, (2000). Starting with the first peak in a cluster (usually the monisotopic peptide), the algorithm generates the theoretical isotopic distribution based on peak height (abundance) and mass value. The computed abundance values are then subtracted from the original peaks at the corresponding m/z values. Following subtraction, any abundance value below a user-specified threshold is set to zero. The step is then repeated, with the remaining (height adjusted) peaks. At each step, the monoisotopic mass value, the original detected abundance and the total abundance are recorded in the final peak list. The deisotoping unit processes previously computed clusters from the dual port RAM (B), writes back partial results in RAM (B) and the final peak list in RAM (A).

When the last peak from the cluster is visited, the total peptide abundance for every monoistope detected is stored back to RAM(A), in the 32 bit area previously used to store the cluster flag $f_{k+1}$. At the same time, the cluster index information (last 12 bits) is overwritten with a flag set to 1 or 0 depending on whether the monoistopic peak at that address is above or below the threshold $\tau_3$.

For example, the monoisotope peaks recovered from the previously clustered fragment displayed in Figure 6 are shown in Figure 7. In this example, two consecutive overlapping monoisotopomers are detected. The abundances of all the higher isotopes of each monoisotope are added to its original monoisotopic abundance. The first monoisotope has its summed abundance of 867.7146 while for the second monoisotope the corresponding overall abundance is 4333.5933. When processing ends, the harvested peak list in RAM(A) is ready to be used to search the protein database.

## 3 RESULTS

The processor was implemented on a FPGA motherboard equipped with a Xilinx Virtex-II XC2V8000 FPGA (8 million gates) and 4 Mbytes ZBT RAM, communicating with the host PC server via a PCI interface (32 bit, 33 MHz).

On the motherboard there are two FPGA devices (Fig. 8). The bigger one (Virtex-II XC2V8000 FPGA) is used to implement user designs—in our case the spectrum processor. The Xilinx Spartan-II FPGA implements the PCI interface between the server PC and the user FPGA from the motherboard. Communication between these two FPGA devices is at 40 MHz on a 32 bits wide data bus. The motherboard has 4 MB of ZBT RAM connected to the user FPGA as shown in Figure 9. This is enough to store 512 K samples of mass-abundance pairs on 32 bits each.

The actual design occupies about 70% of the FPGA's logic resources and 18% of the FPGA's I/O resources. The server is a Dual 3.06 GHz Xeon processor machine with 4 GBytes RAM. The block scheme of the system is given in Figure 8.

The mass spectrum is transferred into the ZBT RAM via the PCI interface in two steps, first the mass and second the abundance data vectors.

The design can be easily ported, however, to a new version of the motherboard that supports PCI-X standard (64 bits at 133 MHz) which will allow transfer of the abundance and mass data streams at the same time.

All arithmetic operations are performed using 32-bit signed fixed-point binary number representation of mass and abundance values, with 12 bits after the radix point.

### 3.1 Spectral processing

The basic steps in processing a mass spectrum are largely the same, irrespective of the software used: smoothing, baseline subtraction and centroiding/deisotoping. For the FPGA based approach to be useful, the quality of the processed spectra should be at least as acceptable as those processed by software. A recombinant protein designed as an internal standard for multiplexed absolute protein quantification (Beynon *et al.*, 2005; Pratt *et al.*, 2006) was digested with trypsin to release

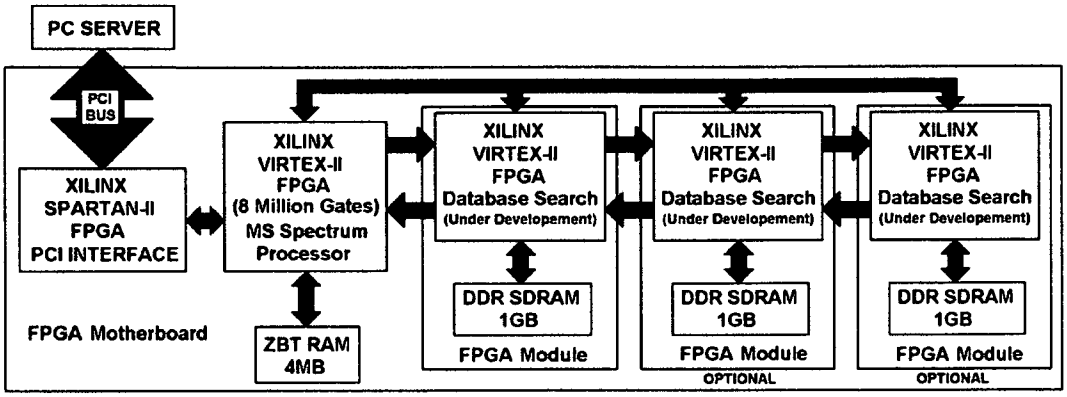



Fig. 8. Block scheme of the system.

20 limit peptides of known identity. Digested material was analysed by MALDI-ToF MS and raw data was processed separately using MassLynx, a commercial mass spectrometry software, and FPGA.

Data was processed using MassLynx software to remove background noise using polynomial order 10 with 40% of the data points below this polynomial curve and a tolerance of 0.01. Spectral data was also smoothed by performing two mean smooth operations with a window of three channels. The processed spectra were compared as a scatterplot of the centroid intensity values (relative to base peak) for data analysed in each way. The centroided spectra are highly comparable and the FPGA identifies the same peaks as the commercial product (Masslynx). Moreover, the intensities of the different peaks correlated well, irrespective of the method used to process the spectrum (Fig. 10) identified by the software.

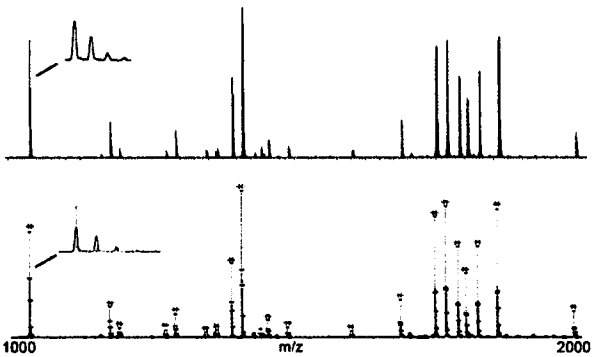



Fig. 9. A recombinant protein designed as an internal standard for multiplexed absolute protein quantification (Beynon *et al.*, 2005; Pratt *et al.*, 2006) was digested with trypsin to release 20 limit peptides of known identity. Digested material was analysed by MALDI-ToF MS and raw data was processed separately using MassLynx software and FPGA.

### 3.2 Deconvolution

A valuable test of effective deisotoping is provided by the resolution of isotopomer distributions derived from asparagine containing peptides and the deamidated cognate peptide. If a peptide contains the sequence Asn-Gly in particular there is a marked propensity for this to be converted non-enzymically to Asp-Gly, with the result that the deamidated peptide is 1Da heavier ($-NH_2$ to $-OH$). We tested this part of the analysis using MALDI-ToF peptide mass spectra derived from in-gel digestion of glyceraldehyde 3 phosphate dehydrogenase. A peptide generated by tryptic digestion has the sequence VKVGVNGFGR (monoisotopic mass 1031.59 Da, creating a singly charged ion $[M+H]^+$ of 1032.59 m/z) which is readily deamidated to VKVGVDGFGR (monoisotopic mass 1032.57 Da, creating a singly charged ion $[M+H]^+$ of 1033.57 m/z). To assess the ability of the FPGA implementation to deconvolute complex and overlapping spectral data, we generated a set of spectra for this peptide. The 1Da mass shift on deamidation generates a series of mass spectra that are strongly overlapping (Fig. 11).

Previously, we have assessed the proportion of acid and amide by a non-linear least squares iterative curve fitting procedure that explains the observed mass spectrum by optimizing the proportion of acid and amide variants, from theoretical spectra for the acid and amide species generated using the Protein Prospector MSIsotope tool (http://prospector.ucsf.edu/ucsfhtml4.0/msiso.htm). The correlation between the calculation of acid:amide proportion was exactly the same, irrespective of whether the FPGA implementation or the non-linear least squares method was used (Fig. 12). Thus, the hardware solution was able to deconvolute overlapping spectra with ease and yield the same results as previous methods.

### 3.3 Speed gains

The impact of the spectral length on processing time was measured using spectra with various lengths but constant isotopic composition and noise levels. The reference design was compiled in C and was simulated on a dual processor server using 3.06 GHz Xeon devices running Windows XP Professional operating system. The FPGA processor had an internal clock frequency of 180 MHz. In order to evaluate how the number of

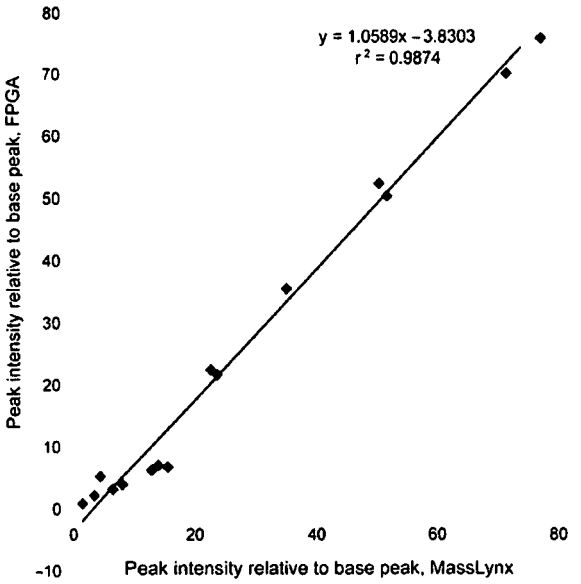



**Fig. 10.** Measurement of peak intensity by FPGA and commercial software.

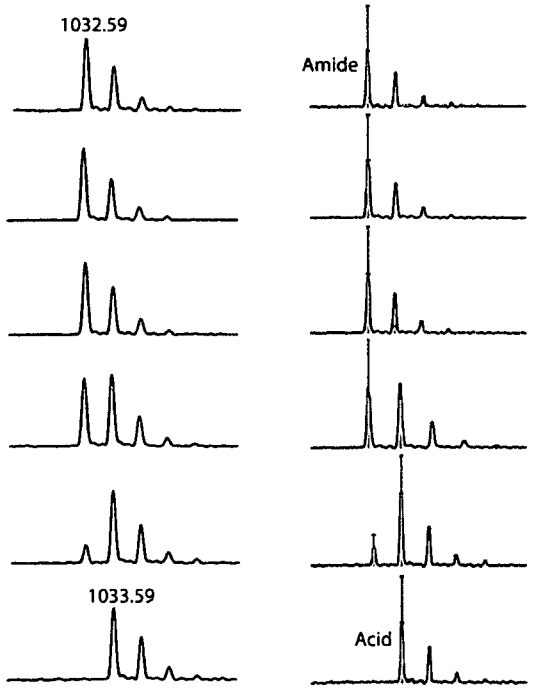



**Fig. 11.** Test data for deconvolution of mass spectra. A series of spectra were acquired for a peptide that undergoes deamidation using MALDI-ToF mass spectrometry (left hand column). After processing the data with the FPGA implementation, the proportions of acid ([M + H]+ = 1032.59 m/z and amide ([M + H]+ = 1033.59 m/z) were calculated, and indicated on the processed spectra by vertical drop lines, headed by asterisks (right hand column).

data points in a spectrum relates to speed gain, spectra with different number of mass-abundance pairs were processed. The software processing routine was repeated 30 times for each mass spectrum data set and timed. The average time was used to calculate speed gains. It should be mentioned that the software processing time does not account for data transfers and memory initialization operations. Only the main computational loop was timed. Initializations for example, add on average 30 ms to the C processing time. The results are summarized in Table 1.

The average speed gain for processing spectra of different lengths is 122. Of course, implementations in instrument manufacturers' software are somewhat slower, and spectral processing such as obtained here can take several tens of seconds.

It is interesting to note that on a single processor server having the same configuration as the dual processor—except of the number of processors—the average time of processing the largest spectrum of 200 976 mass-abundance pairs was 204.71 ms which corresponds to a speed gain of about 180.

Processing time is less dependent on the number of signal peaks in the mass spectrum. Although, clustering and deisotoping processes are time consuming and depend on the spectral composition (i.e. the higher the isotopic abundance, the larger the number of iterations that have to be performed), the number of monoisotopes and their isotopic contributions is far less then the entire spectrum data. As a consequence, peak identification, which involves processing the entire spectrum, represents the most time consuming operation, giving the bulk of the total processing time.

## 4 DISCUSSION

We have successfully demonstrated that processing of a mass spectrum can be very effectively implemented as a hardware solution in a high-density FPGA. The performance is comparable in terms of quality of the processed spectrum, and spectra can be processed at much higher rates than obtained through software alone. For example, our FPGA implementation of the PMF algorithm can process in 1 s over 900 mass spectra consisting of 200 000 mass-abundance pairs. When implemented alongside a hardware implemented database search algorithm, which should deliver a match in <100 ms, the goal of real-time peptide mass fingerprinting seems eminently achievable. Another very exciting prospect is that FPGAs will enable the fast execution of 'intelligent' optimization protocols of instrument settings and spectrum processing, which take prohibitively long time to run even on a high-end workstation. For example, the closed-loop multi-objective optimization approach proposed recently by O'Hagan *et al.* (2005), which employs Genetic Algorithms and Genetic Programming to determine optimal instrument settings and remove noise, reportedly takes from 20 min and up to 118 h to run.

The motherboard can be configured to have up to three additional FPGA modules that can be plugged into dedicated motherboard slots. These modules will be used to implement the database search. Each FPGA module has one Virtex-II XC2V8000 FPGA device and 1GB of DDR SDRAM that can

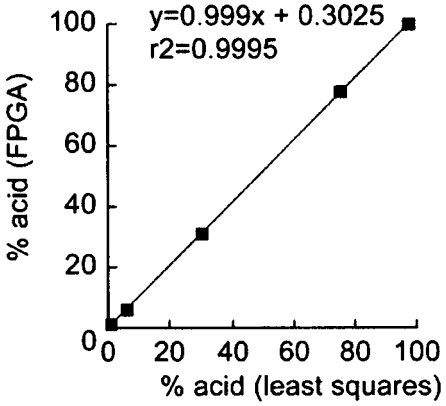



**Fig. 12.** Performance of the FPGA implementation in spectra deconvolution. The set of mass spectra described in Figure 7 were used as source data for the calculation of the proportion of acid and amide whether assessed by a least squares method or using the FPGA implementation.

**Table 1.** Benchmark results of the spectrum processor implementations

| Spectrum size | Timing [ms]: Dual Xeon 3GHz processors | Timing [ms]: Virtex-II FPGA, 180MHz clock | Speed gain |
|---|---|---|---|
| 25 488 | 20.27 | 0.1632 | 124.20 |
| 50 448 | 31.23 | 0.3105 | 100.56 |
| 75 168 | 47.33 | 0.4557 | 103.86 |
| 101 040 | 62.50 | 0.5607 | 111.46 |
| 125 184 | 79.17 | 0.7557 | 104.76 |
| 150 114 | 114.33 | 0.8547 | 133.76 |
| 175 104 | 130.20 | 1.0024 | 129.88 |
| 200 976 | 188.63 | 1.1219 | 168.13 |

easily hold the entire protein database. Each module is connected with the motherboard user FPGA implementing the spectrum processor and with other two modules via a 64 bit, 66 MHz local bus. This architecture will enable the implementation of parallel searches at FPGA level as well as across modules.

There are three types of proteomics: identification proteomics, characterization proteomics and quantitative proteomics. The core technology in many of these applications is mass spectrometry, but power of modern instrumentation brings with it the penalty of highly information-rich data streams at very high rates. As such, the bottleneck is moving from data acquisition to data processing. In identification proteomics and in quantification proteomics in particular, hardware solutions such as described here, could solve this bottleneck, and increase proteomics throughput considerably.

## ACKNOWLEDGEMENTS

This work was supported by BBSRC. The authors gratefully acknowledge the support of Xilinx Inc. who donated the devices and design tools used in this study.

*Conflict of Interest*: none declared.

## REFERENCES

Anish,T.A. *et al.* (2005) Hardware-accelerated protein identification for mass spectrometry. *Rapid Commun. Mass Spectr.*, **19**, 833–837.

Beynon,R.J. *et al.* (2005) Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides rates. *Nat. Methods*, **2**, 587–589.

Breen,E.J. *et al.* (2000) Automatic Poisson peak harvesting for high throughput protein identification. *Electrophoresis*, **21**, 2243–2251.

Fagin,B. *et al.* (1993) A special-purpose processor for gene sequence analysis. *Comput. Appl. BioSci.*, **9**, 221–226.

Guccione, A.S. and Keller, E. (2002) Gene Matching Using Jbits, Proceedings of the Reconfigurable Computing Is Going Mainstream, 12th International Conference on Field-Programmable Logic and Applications, 1168–1171.

Guerdoux-Jamet,P. and Lavenier,D. (1997) SAMBA: hardware accelerator for biological sequence comparison,. *Comput. Appl. BioSci.*, **13**, 609–615.

Hughey,R. (1996) Parallel hardware for sequence comparison and alignment. *Comput. Appl. BioSci.*, **12**, 473–479.

Lavenier,D. (1998) Speeding up genome computations with systolic accelerator. *SIAM News*, **31**, 1–8.

Marongiu,A. *et al.* (2003) Designing hardware for protein sequence analysis. *Bioinformatics*, **19**, 1739–1740.

O'Hagan,S. *et al.* (2005) Closed-Loop, Multiobjective Optimization of Analytical Instrumentation: Gas Chromatography/Time-of-Flight Mass Spectrometry of the Metabolomes of Human Serum and Yeast Fermentation. *Analytical Chem.*, **77**, 290–303.

Oliver,T. *et al.* (2005) Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. *Bioinformatics*, **21**, 3431–3432.

Pratt,J.M. *et al.* (2006) Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protocols*, **1**, 1029–1043.

Samuelsson,J. *et al.* (2004) Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting,. *Bioinformatics*, **20**, 3628–3635.

Savitzky,A. and Golay,M.J.E. (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chem.*, **36**, 1627–1639.

Simmler,H. *et al.* (2004) Real-Time Primer Design for DNA Chips. *Interscience Concurr. Comput.: Pract. Exper.*, **16**, 855–872.

Wozniak,A. (1997) Using video-oriented instructions to speed up sequence comparison. *Comput. Appl. BioSci.*, **13**, 145–150.

XILINX (2004) Distributed Arithmetic FIR Filter V9.0, DS240, Xilinx Inc.

# Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes

Julie M Pratt[1], Deborah M Simpson[1], Mary K Doherty[1], Jenny Rivers[1], Simon J Gaskell[2] & Robert J Beynon[1]

[1]Department of Veterinary Preclinical Sciences, University of Liverpool, Crown Street, Liverpool, L69 7ZJ, UK. [2]Michael Barber Centre for Mass Spectrometry, School of Chemistry, University of Manchester, Manchester M13 9PL, UK. Correspondence should be addressed to R.B. (r.beynon@liv.ac.uk), website URL http://www.liv.ac.uk/pfg.

Published online 17 August 2006; doi:10.1038/nprot.2006.129

**An important area of proteomics involves the need for quantification, whether relative or absolute. Many methods now exist for relative quantification, but to support biomarker proteomics and systems biology, absolute quantification rather than relative quantification is required. Absolute quantification usually involves the concomitant mass spectrometric determination of signature proteotypic peptides and stable isotope-labeled analogs. However, the availability of standard labeled signature peptides in accurately known amounts is a limitation to the widespread adoption of this approach. We describe the design and synthesis of artificial QconCAT proteins that are concatamers of tryptic peptides for several proteins. This protocol details the methods for the design, expression, labeling, purification, characterization and use of the QconCATs in the absolute quantification of complex protein mixtures. The total time required to complete this protocol (from the receipt of the QconCAT expression plasmid to the absolute quantification of the set of proteins encoded by the QconCAT protein in an analyte sample) is ~29 d.**



## INTRODUCTION

Most proteomics studies to date have delivered relative quantification, expressing the changes in the amount of proteins in the context of a second cellular state or control sample[1–3]. Such studies do not facilitate the generation of large databases of results, with data not being transferable between different laboratories.

If proteomics is to support the emergent fields of protein biomarker discovery (whether in medical diagnostics or toxicology for drug discovery) or to provide the rigorous data needed for systems biology, absolute quantification is needed. Absolute quantification relies on the well-established precepts of stable isotope dilution, specifically the use of labeled peptide internal standards that are characterized and quantified by mass spectrometry (MS)[4,5]. These internal standards are currently synthesized *de novo* by chemical methods. This requires the individual synthesis in labeled form, purification and quantification, of each peptide for use as an internal standard. Complex studies would require the synthesis of large numbers of peptides at significant cost, and each would have to be quantified individually.

We describe here a method for the design, expression and use of artificial proteins (QconCATs) that are concatamers of Q peptides, generated by chemical or endoproteolytic cleavage, for a group of proteins under study[6]. The QconCAT proteins are expressed in *Escherichia coli* and are readily labeled with stable isotopes by growth in the presence of stable isotope–labeled precursors. The labeled QconCAT proteins are then purified, quantified and added to complex protein mixtures in known amounts. Endoproteolytic (and/or chemical) fragmentation of the QconCAT–analyte mix releases each of the QconCAT peptides in a strict stoichiometry of 1:1, and MS analysis allows the quantification of each represented peptide of the analyte (see **Fig. 1** for a schematic of the overall process). In the specific protocol described here we focus on the use of trypsin as the endoproteinase, but other endoproteinases and/or chemical cleavage could be used as an alternative method of generating suitably sized peptides from the analyte and identical peptides, to act as internal standards, from the labeled QconCAT protein.

Unlabeled QconCAT proteins could also provide the basis for absolute quantification if differential isotope labeling via derivatization of proteins or peptides is incorporated in the analytical procedure. Use of the same QconCAT protein for each quantification experiment allows direct comparison of results between each experiment and between laboratories, and will facilitate the construction of the large data sets needed for toxicological and diagnostic studies in which even control samples vary considerably. We will also describe a refinement to the technique that even allows the quantification of the QconCAT standard between laboratories. Because the quantification data are in absolute terms—expressed, for example, as picomoles of protein per gram of tissue or per number of cells—knowledge of the number of molecules of a protein present per cell in a given state will underpin the generation of testable mathematical models in systems biology.

## APPLICATIONS

QconCATs can be applied to the absolute quantification of any peptide that can be reproducibly generated by endoproteolytic or chemical cleavage from any sample source. Here we discuss only a sampling of the potential applications.

### General

- In comparative proteomics the shift from relative to absolute quantification permits comparison of results, not only between different cellular states within an experiment, but also between different experiments and different laboratories.
- An unlabeled QconCAT could be subjected to labeling *in vitro* using one of the many reagents that have been advocated for comparative proteomics, enhancing all of these technologies to absolute quantification.

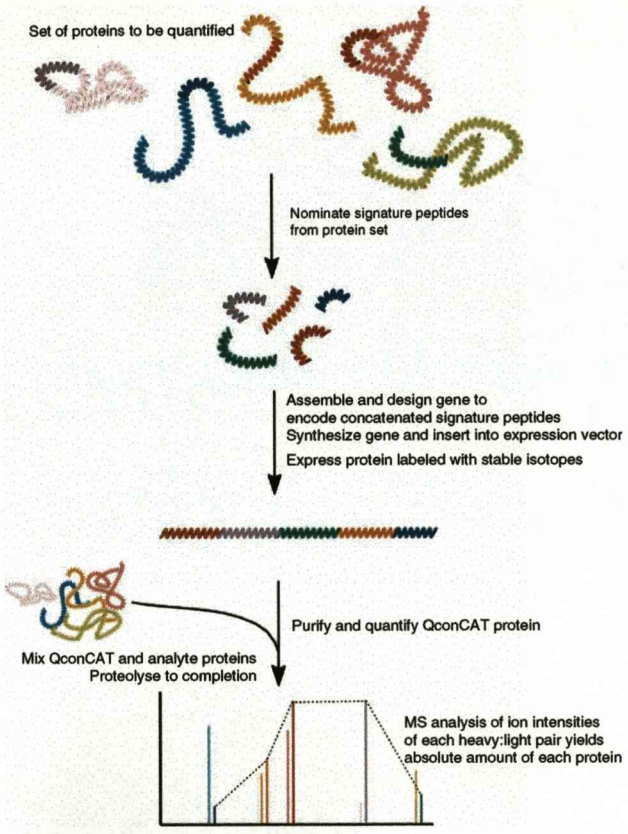



**Figure 1** | General principle of a QconCAT quantification experiment.

- Systems biology requires absolute quantification, including the definition of processes in molecular terms and the generation of data to allow mathematical modeling.
- Determination of the absolute stoichiometry of components in subcellular particles requires absolute quantification.
- Analysis of isoform expression patterns (e.g., cytochrome P450s) is possible using this process. Each QconCAT peptide would report on a peptide that contains the variant amino acid(s), permitting one to distinguish among several variants.
- QconCATs could be used to quantify many or all proteins expressed in a control strain or cell culture grown under standardized conditions. Subsequently, the same strain or cell culture would be labeled by growth in the presence of stable isotope precursors under the same conditions. These fully labeled cells would then be used as quantification standards in comparative studies. The indirect quantification approach has the advantage that, for the second quantification phase, any labeled peptide (not only those embedded in QconCATs) could be used, providing the opportunity for alternative enrichment and quantification approaches.

**Medical**

- Diagnostic/protein biomarker validation and analysis could be facilitated by QconCATs for accurately measuring changes in the protein components of body fluids and thereby adding another dimension to computer training sets for identifying patterns indicative of disease.
- For routine diagnostic tests, QconCATs could allow determination of levels of key biomarkers in clinical samples including blood, urine, cerebrospinal fluid (CSF), synovial fluid and bronchoalveolar lavage.
- QconCATs could allow the monitoring of changes in protein levels in response to exposure to drugs, as well as identifying changes relating to toxicology and routine monitoring of protein biomarkers to accelerate the process of drug discovery.

**Agricultural**

This technology could be used in routine determination of levels of proteinaceous contamination from other sources in foodstuffs, for example, signatures from genetically modified sources. It would also facilitate monitoring of the effect of pesticides and herbicides.

**Protein chemistry**

QconCATs introduce the concept of 'designer protein' to proteomics. Specific QconCAT proteins could therefore be used to assess the properties of defined peptides in MS and MS/MS analysis.

**Post-translational modifications**

The use of QconCATs to study some post-translational modifications is limited, because only the unmodified analyte peptide can be quantified using a Q-peptide. Nevertheless, we can envisage two methods of using QconCATs to provide a preliminary study of post-translational modifications. One approach involves selecting a minimum of two peptides for each protein of interest; one peptide is not modified and the other is modifiable. Comparison of the quantification of the two peptides should allow the determination of the amount of modified peptide present. The second approach is to select the modified peptide as the Q-peptide to report on the protein and carry out the quantification before and after treatment of the sample to remove post-translational modifications—for example, alkaline phosphatase to hydrolyze phosphopeptides. Quantification before and after treatment will reveal the quantity of protein in the modified state. These methods are not, however, immediately applicable to the unraveling of more complex patterns of post-translational modification, in which several sites of modification are possible on the same peptide.

For some modifications, such as the primary sequence change elicited by limited proteolysis (such as apoptosis), it might be argued that QconCAT-type approaches are the only ways to accurately quantify the proteolysed and unproteolysed variants of the target analyte.

## EXPERIMENTAL DESIGN

There are two stages in a QconCAT experiment: the design phase and the implementation phase leading to the construction of a QconCAT (**Fig. 2**). Once the design exists as a DNA sequence it can then be inserted into a suitable vector, transferred to a suitable expression system, expressed, labeled and used in quantification experiments. The first stage reflects the analytical and decision-making processes that encompass the nomination of candidate proteins and peptides, followed by their realization as a codon-optimized sequence in which opportunities for stable RNA secondary structure formation are minimized. A Gantt chart outlining the timeline for the procedure is given in **Figure 3**.

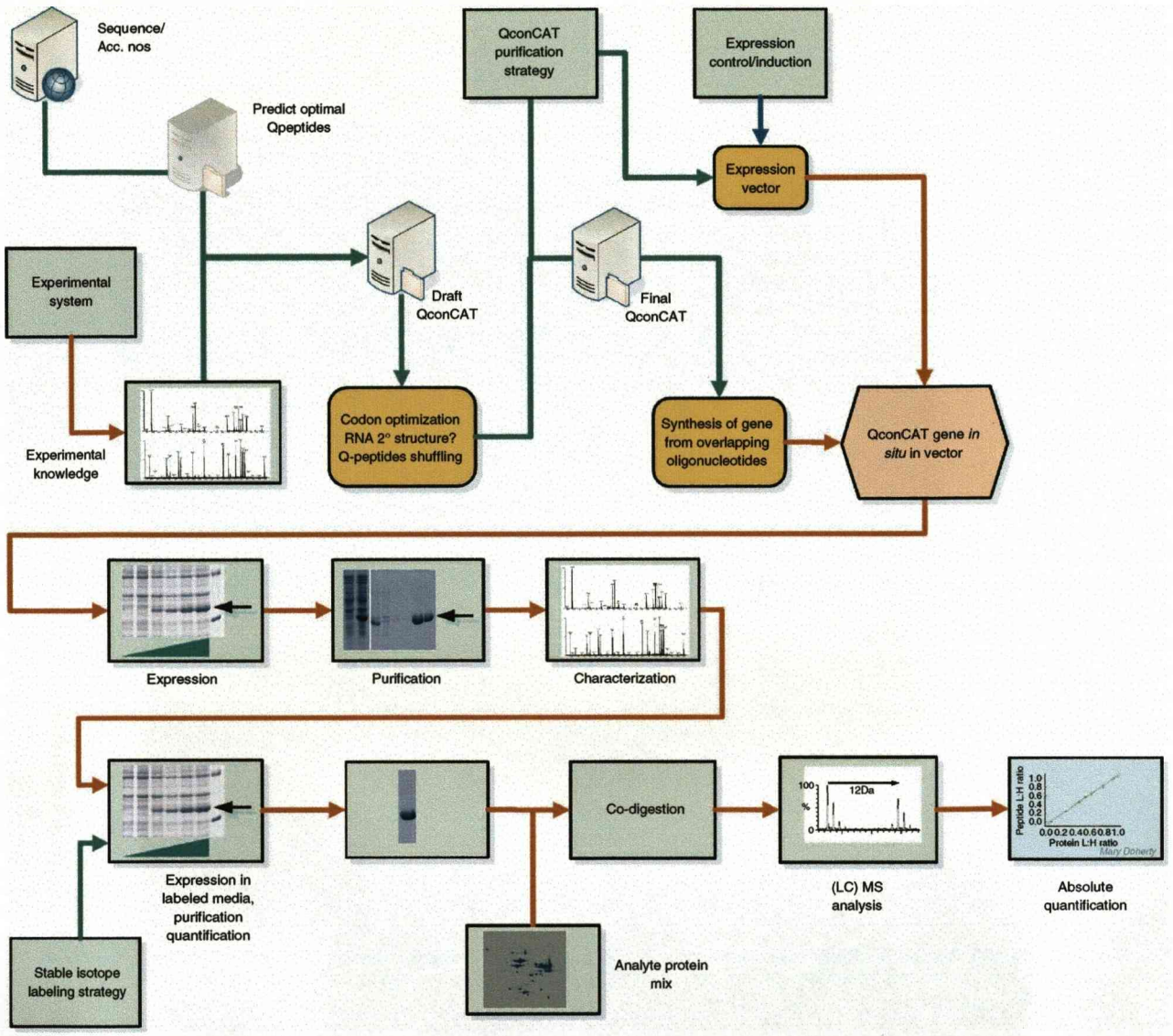



**Figure 2** | Overall workflow for QconCAT quantification. A QconCAT experiment consists of the design phase and implementation phase.



### Design of QconCAT genes and construction of plasmids encoding QconCAT proteins

To quantify the group of proteins, a unique peptide (a Q-peptide) must be selected as a surrogate 'signature' peptide for each protein to be quantified. This selection depends on the inherent properties of the peptide and the cleavage method chosen for the analysis of the analyte. There are no restrictions on the protein-containing samples that can be quantified using QconCAT technology. However, in the absence of complete genome information, full sequencing of the signature peptide (probably by tandem MS) would be required before incorporation of this peptide into a designed QconCAT.

### Selection of cleavage method

Most proteomics studies have used trypsin as the method of cleavage of proteins into peptides suitable for MS analysis. Trypsin has the advantage of a rigorously expressed sequence specificity (cleaves C-terminal to arginyl and lysyl residues, except arginyl-prolyl and lysyl-prolyl sequences, which are not cleaved) and minimal autoproteolysis (recombinant trypsin). For many proteins cleavage with trypsin generates a good distribution of peptides within the range 600–4,000 Da, ideal for analysis by MS. Because only one peptide is required in a QconCAT protein to represent each protein under study, if trypsin does generate a suitably sized peptide, with the additional properties described in the next section, trypsin will be the endoproteinase of choice. Moreover, tryptic peptides readily generate doubly charged ions ($[M+2H]^+$) that extend the usefulness from MS to MS/MS analysis and increase the scope of quantification to include monitoring of single or multiple reactions. However, for integral membrane proteins, with several membrane-spanning regions, there is an under-representation of lysyl and arginyl residues. In addition, cleavage sites may be

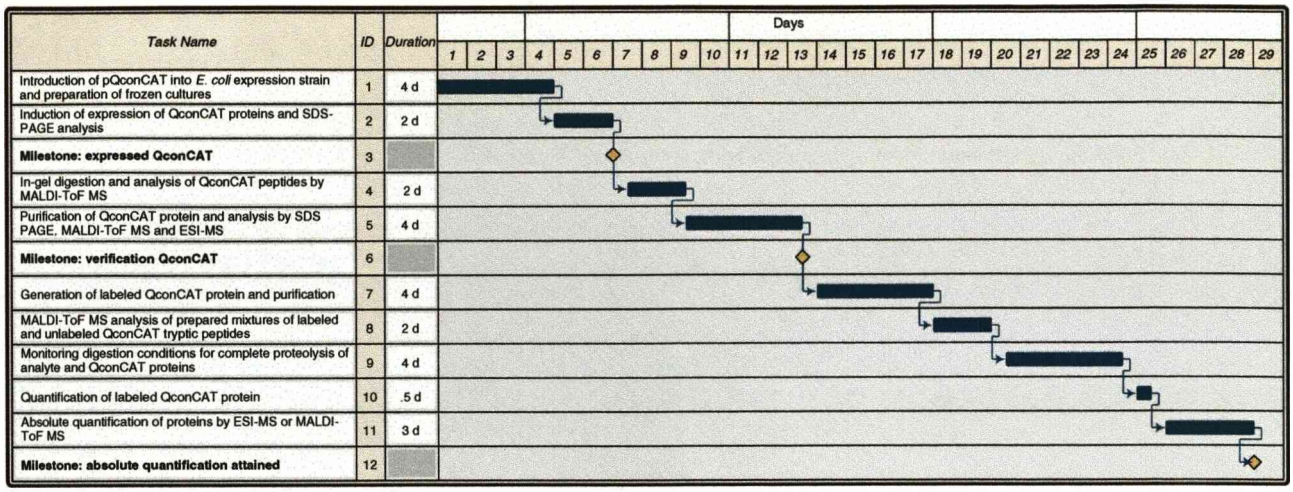



**Figure 3** | Gantt chart for a typical QconCAT experiment.

inaccessible to endoproteinases because of steric hindrance from the lipid bilayer or localization of cleavable sequences within the lumen of membrane vesicles. More efficient study of these proteins requires alternative solubilization[7,8] and cleavage strategies[9]. Methods that have been developed include (i) chemical cleavage at methionine residues using cyanogen bromide[7], in combination with trypsin[9] and (ii) chymotrypsin (which cleaves C-terminally at phenylalanyl, tyrosyl and tryptophanyl residues), and staphylococcal peptidase I (which cleaves at glutamyl residues), either in combination or singly[9]. However, for some of these peptidases, the specificity is sufficiently relaxed that preliminary experiments will be needed to ensure exclusive release of a suitable peptide (with no overlapping or alternative cleavages). The cleavage strategy should therefore be decided and tested for the group of proteins to be quantified before QconCATs are designed, so that the appropriate peptides can be selected for inclusion to match the planned analyte protein fragmentation strategy. A relaxed specificity can be tolerated provided that the same peptide is generated quantitatively from both QconCAT and analyte; in practice this might be difficult to achieve.

### Selection of Q-peptides

Q-peptides are selected for uniqueness of mass, propensity to ionize and detectability in MS, the presence of specific amino acid residues required for labeling and the absence of amino acid residues that can cause problems (cysteine, methionine) (see http://www.qconcat.com for additional guidelines). The propensity of a selected peptide to ionize is often known, because the sample has been analyzed by MS previously. It is not yet possible to predict ion intensities using knowledge-based approaches, but this may make peptide selection easier in the future. At the moment we favor a pragmatic solution. The inclusion of two Q-peptides per protein might reduce the risk associated with poorly ionizing or otherwise difficult peptides. In addition, an improvement in ion intensity of peptides lacking arginine, but carrying lysine, can be achieved by guanidination, and this can be included as a relatively simple step before analysis by MS[10–12].

In some instances it will be impossible to avoid inclusion of peptides containing cysteinyl and methionyl residues, for example. For these QconCATs it is essential to be aware of the need to maintain conditions that ensure that the oxidation state of the Q-peptide is the same as the analyte, because any discrepancy between the two would compromise the quantification. Finally, for those groups that have been using individual chemically synthesized peptides, for comparison of quantifications using QconCATs with chemically synthesized stable isotope-labeled peptides, a peptide available in purified form could be included in the QconCAT sequence and its behavior as part of a QconCAT compared to that when present in an analyte mixture, as a purified, quantified peptide.

### Construction of QconCAT genes

The Q-peptide sequences once selected are randomly concatenated *in silico* and used to direct the design of a gene, codon-optimized for expression in *E. coli*. The predicted transcript is analyzed for RNA secondary structure that might diminish expression, and if undesirable secondary structure features are present, the order of the peptides is altered. N-and C- terminal sequences are added as sacrificial structures, protecting the assembly of true Q-peptides from exoproteolytic attack during expression. Additional peptide sequences are added to provide an initiator methionine and, if appropriate, a single cysteinyl residue as an alternative means of quantification. The QconCAT genes are synthesized by companies (e.g., Entelechon GmbH) who specialize in gene synthesis for the generation of recombinant proteins. The construction and synthesis of the gene to encode a QconCAT is not described here (see **Supplementary Methods** in ref. 6), and the protocol assumes that the experimenter is starting with a QconCAT plasmid.

### Expression and purification of QconCAT proteins

Once an expression plasmid encoding the QconCAT protein has been generated (pQconCAT), the plasmid is introduced into an appropriate *E. coli* expression strain. A single transformant is then grown in rich medium and the expression of the QconCAT protein is induced and analyzed by SDS-PAGE; finally, the QconCAT band is subjected to in-gel proteolytic digestion and analysis by MS. This simple characterization allows rapid confirmation of adequate QconCAT protein expression and also allows an early evaluation of the signal intensities of the chosen Q-peptides. These preliminary characterizations of the QconCATs are advisable, before

proceeding to labeling with relatively expensive heavy isotope–labeled precursors.

Because QconCAT proteins are concatenations of short peptides (typically 10–20 amino acids) from different proteins, it is difficult to predict their properties. However, it might be anticipated that expression and purification of QconCAT proteins should be straightforward, because these proteins do not have to be functionally folded. Furthermore, they can be purified and used in denatured form, because the only requirement is that they be susceptible to digestion by endoproteolytic enzymes (trypsin, endopeptidase ArgC, endopeptidase LysC, chymotrypsin, endopeptidase Asp-N, staphylococcal peptidase I) or by chemical cleavage. Therefore, expression of a QconCAT does not bring the complexities that are sometimes encountered when a protein must be heterologously expressed and folded in the correct form.

**Stable isotope–labeling strategy**

Labeled QconCAT proteins are generated by heterologous expression, most simply by growth of the *E. coli* host carrying the pQconCAT in defined medium supplemented with an appropriate stable isotope–labeled precursor. An alternative method would be to use QconCAT expression plasmid DNA to prime a cell-free coupled transcription/translation system, supplemented with the chosen heavy isotope precursor(s). Generation of QconCAT proteins in the latter system may be more compromised in terms of sequence fidelity and also gives lower yields. However, constructs that are toxic to *E. coli* or are extremely unstable (as a result of proteolysis) may be expressed *in vitro* when they cannot be expressed in *E. coli*.

Uniform labeling with $^{13}C$ or $^{15}N$ would ensure that every QconCAT peptide is comprehensively labeled; however, each labeled (heavy) peptide will be separated in mass by differing amounts from its unlabeled (light) counterpart, complicating the mass spectroscopic analysis. If Q-peptides are selected that contain at least one instance of a specific amino acid, that amino acid can be incorporated in stable isotope–labeled form. Perhaps the preferred strategy reflects that because most QconCAT proteins are assemblies of tryptic peptides, incorporation of [$^{13}C_6$]lysine and [$^{13}C_6$]arginine would ensure that most Q-peptides would be singly labeled and the mass offset between heavy and light peptides would be a constant 6 Da. Q-peptides that contain internal Arg-Pro or Lys-Pro sequences will of course contain two instances of the labeled amino acid, yielding a mass offset of 12 Da. We advocate the use of $^{13}C$-labeled amino acids instead of $^{2}H$-labeled amino acids, because of the difference in elution time of peptides containing $^{2}H$-labeled amino acids in liquid chromatography and the probable decrease in quantification accuracy caused[13]. Alternatively, unlabeled QconCATs could be labeled *in vitro* using one of the many reagents that have been advocated for comparative proteomics.

**Extent of proteolysis**

The next stage is to determine the conditions for complete digestion of both the analyte and QconCAT proteins. First, a true internal standard will, apart from the stable isotope labeling, be otherwise exactly the same as the analyte. In the QconCATs, this is clearly not the case, because each surrogate peptide in the concatamer is in a sequence context that is different from the same peptide in the analyte protein. This means that the sequence contexts of the scissile bonds differ, and rates of hydrolysis could therefore be different for the QconCAT and the analyte proteins. Our experience with QconCATs thus far is that they lack higher order structure and are readily and rapidly digested by endopeptidases such as trypsin or endopeptidase LysC. Proteolytic digestion of the analyte proteins, on the other hand, is mostly impeded by the presence of higher order structure, which restricts access to proteases[14–18]. However, provided that the analyte proteins are extensively denatured and that the reaction is allowed sufficient time to proceed to completion, this need not be a serious issue. We have used the following methods to denature protein mixtures in anticipation of proteolysis for quantification.

- Precipitation with an acid such as 10% (wt/vol) trichloroacetic acid. The ether-washed pellet (analyte plus QconCAT) can then be proteolysed back into solution.
- Heat treatment (95 °C for 10 min). Again, precipitated material can be resolubilized by proteolysis.
- Treatment with a chaotrope such as urea or guanidinium hydrochloride, which would then have to be removed before proteolysis.

Acetone precipitation may not be a sufficiently potent denaturation step to enhance proteolysis of many proteins—indeed, it has been used as a precipitation step in protein purification protocols where activity is preserved.

If the analyte proteins contain disulfide bonds, it is also necessary to reduce and alkylate cysteinyl residues. If the QconCAT must contain cysteinyl residues as well, then the analyte-QconCAT should be mixed before reduction/alkylation.

The rapid digestion (usually within 1 min at typical substrate-to-protease ratios of 50:1 for trypsin) of the QconCAT means that it can also be used as a reference in pilot studies designed to assess the degree of proteolysis of the analyte mixture. We would stress that the requirement for complete digestion has rarely been specifically addressed in any quantification studies, and that this step is critical to any such experiment, whether relative or absolute. Even single-protein quantification by peptide chemical synthesis requires complete hydrolysis of the analyte. For identification proteomics, complete digestion may not be required, and the extent of digestion may not therefore be routinely assessed, but we must emphasize that incomplete digestion means incorrect quantification.

**Quantification of the QconCAT**

Because the absolute quantification strategy involves relative quantification against an accurately determined internal standard, it is worth emphasizing the principles that are involved.

Absolute quantification with a QconCAT is only as good as the quantification of the QconCAT itself. The first step of any QconCAT experiment is thus the determination of the concentration of His-tag purified QconCAT. Several approaches are possible.

- Determination of a QconCAT by protein assay. It is possible to determine the QconCAT concentration by protein assay, and this will provide an acceptable degree of accuracy for many purposes. However, different methods of protein assay yield different results. Furthermore, it is not possible to use the exact protein as a reference, and often a protein such as BSA must be used. Of the different protein assays available, the biuret method, which is insensitive to the nature of the protein being assayed, is preferred. This method can consume a lot of protein (1 mg for a 1-ml assay), but microbiuret methods are available. Moreover, new spectrophotometric

instruments are capable of operating on vanishingly small volumes, and thus sample consumption might be less of a problem.

- Determination of QconCAT by densitometry. It is possible to run the QconCAT on a one-dimensional SDS-PAGE gel and, in parallel lanes, to include known amounts of a protein standard. After staining, densitometry of the lanes allows interpolation of the amount of the QconCAT relative to the standard; thus, its concentration can be assessed. However, this method also suffers from different response factors to staining protocols.
- Determination of QconCAT by Kjeldahl assay. This approach measures total nitrogen by a complex and fairly hazardous process, and although it probably delivers excellent results, its use is not warranted.
- Determination of QconCAT by amino acid analysis. Complete acid hydrolysis (6 M HCl, 110 °C, oxygen-free environment) of a QconCAT would release free amino acids, which can then be assessed by any one of the many methods for amino acid analysis. In nearly every method the amino acid content of the QconCAT will itself be assessed relative to a set of standards, which must obviously be carefully prepared. However, acid hydrolysis affects many amino acids: valine and isoleucine bonds are less easily hydrolyzed, threonine and serine are partially destroyed, methionine can be oxidized during acid hydrolysis, asparagine and glutamine are both converted to the respective acidic residue, and tryptophan is completely destroyed.
- Assay based on specific amino acids in the QconCAT. There is a series of assays that could be used to determine the concentration of specific amino acids in the intact QconCAT. For example, a reagent reactive to primary amino groups (lysine and the α-amino group of the N terminus) such as fluorescamine or *o*-phthaldialdehyde-*N*-acetylcysteine could be used. The advantage of this approach is that fluorescence-based assays are sensitive, there will usually be several reaction sites in a QconCAT and the fluorescence yield of all primary amines is very similar—meaning that a simple amine reagent can be used as a standard.

Colorimetric assays for specific amino acids in intact proteins are not common. We have previously included a single cysteine residue in a QconCAT to facilitate determination by Ellman's reagent (DTNB, dithio *bis*-2-nitrobenzoic acid). However, the extinction coefficient of DTNB is 13,600 $M^{-1}$ $cm^{-1}$ at 412 nm, thus, it cannot be used convincingly with a QconCAT concentration of $<0.01$ mM ($A_{412} = 0.136$). A typical QconCAT (molecular weight ~40 kDa) would, at 0.01 mM, be 0.4 mg $ml^{-1}$, and this method therefore consumes a lot of QconCAT.

- Assay based on a common synthetic peptide. An alternative strategy for quantification of QconCATs would be the purchase of an accurately quantified synthetic peptide (in a less expensive unlabeled form, because we are interested in quantifying the labeled QconCAT). Again, we stress that this quantification is only as good as the quantification of the synthetic peptide. However, it would be possible to design a common peptide into every QconCAT and to quantify every protein relative to the same synthetic peptide using MS. For example, the His-tag peptide in the construct described here gives a good mass-spectrometric signal and could be made common to every artificial protein. A common quantification protocol could then be widely disseminated and diminish problems of different peptide or protein quantification.

The quantification of any protein is not without problems, and QconCAT technology brings the same challenges. However, it is essential to recognize the importance of quantification of the QconCAT in the correct analytical context. The QconCAT method has, for any one peptide, an experimental coefficient of variance of ~2%, and this therefore sets the limit on the precision that is available for quantification. However, in many experiments the biological variance will be substantially greater than any analytical step, and again, this defines the quality of the protein quantification that is required. It might be argued that the goal of any analytical method is to make the analytical variance sufficiently small, relative to biological variance, that the emphasis should always be on biological variation. Thus biological replicates are prioritised over technical replicates.

We would venture to suggest that for many experiments a simple colorimetric protein assay is acceptable (either dye binding or biuret), but that for extremely high-precision work, determination of several amino acids or total nitrogen might be best. For many experiments such stringent quantification methods might be considered to be unnecessary. Finally, once a QconCAT has been quantified, there is merit in then determining the $A_{280}$ (1%) of the protein, so that future quantification can be based on simple, nondestructive spectrophotometric measurement.

**Comparison between absolute quantification by QconCATs or synthetic peptides**

So far a direct comparison between peptides prepared by chemical synthesis and by QconCAT has not been undertaken, so we can only discuss the probable pros and cons of the two methods. First, if only a small number of proteins (e.g., ≤10 proteins) is to be quantified, the synthetic peptide approach is probably more appropriate and economical. For the absolute quantification of larger numbers of proteins, because each constituent peptide of a QconCAT is assayed simultaneously, the method is far superior to the quantification of several synthetic peptides in independent experiments, although this feature increases the demand for good QconCAT protein quantification. If the proteins in the group under study differ in concentration by orders of magnitude, it is perhaps easier to add varying amounts of internal standard if they are available as synthetic peptides. However, when designing several QconCATs to cover the range of proteins of interest, it is sensible to group together those proteins of similar abundance within the same QconCAT and adjust the amount of each QconCAT added accordingly. One advantage of a QconCAT is that once the gene has been constructed, a range of labeling methods is available (e.g., Arg, Lys labeling, or complete labeling with $^{15}N$), whereas with synthetic peptides the choice of label must be made in advance of the one-shot synthesis.

**Laboratories specializing in proteomics and MS alone**

The protocol described below is a step-by-step walk through the whole process from the design of the QconCAT to its use in the absolute quantification of complex protein mixtures. This necessarily covers a range of different techniques, which may not be within the capabilities of a typical proteomics laboratory. For

those laboratories at which work with genetically manipulated organisms is not permitted or the required expertise is not present, purified, labeled QconCATs can be obtained commercially (http://www.qconcat.com). Once the group of proteins that you wish to quantify has been decided and the set of surrogate peptides elected, a QconCAT gene can be designed, synthesized, expressed in labeled form, characterized and delivered to the client for immediate use in absolute quantification experiments (http://www.qconcat.com).

**Sources of proteinaceous samples**

Potential analyte sources include human, animal, plant (grasses, shrubs, trees, algae, food), microorganisms, body fluids (blood, serum, CSF, bronchial lavage, semen, vaginal secretion, tears, saliva, sweat, sputum and urine), as well as feces, tissues, cells, cell lines, hair, food, soil, water (as from rivers or the ocean) and sewage. In this exemplar protocol we have included the quantification of proteins in the soluble fraction of chicken skeletal muscle during development immediately after hatching.

## MATERIALS

### REAGENTS

- $^{15}NH_4Cl$ (99% atom percent excess) (CIL Inc.)
- L-arginine hydrochloride (U-$^{13}C_6$, 98%; CIL Inc.)
- L-lysine hydrochloride (U-$^{13}C_6$, 98%; CIL Inc.)
- 20 mM phosphate
- 20 mM, 500 mM imidazole
- 6 M guanidinium chloride
- Protease inhibitors (Complete Protease Inhibitors; Roche)
- Coomassie Plus Protein Assay (Pierce)
- *E. coli* BL21(λDE3) (Stratagene; Promega; Genlanatis)
- Plasmid vector pET21a (Novagen, Merck)
- Luria broth (LB; Merck)
- Luria agar (LA)
- HCl
- NaOH
- Ampicillin sodium salt (Sigma cat. no. A9518-5G)
- Isopropyl β-D-thiogalactopyranoside (IPTG) (Sigma cat. no. 16758-1G)
- Glycerol
- Bromophenol blue
- Ammonium bicarbonate solution
- Trypsin, sequencing grade (Roche)
- Formic acid
- Acetonitrile (ACN)
- Iodoacetamide
- Trifluoroacetic acid (TFA)
- α-cyano- 4-hydroxycinnamic acid
- Bugbuster Protein Extraction Reagent (Novagen; EMD Biosciences)
- Lysozyme (from chicken egg white; Sigma cat. no. L-7651)
- HisTrap HP (GE Healthcare UK, Ltd.)
- StrataClean Resin (Stratagene)
- 7 M ammonium hydroxide
- Sodium phosphate buffer, pH 7.0
- $Na_2HPO_4 \cdot 7H_2O$
- $KH_2PO_4$
- $NH_4Cl$
- $MgSO_4$
- Glucose
- Thiamine
- $CaCl_2$
- Amino acids

### REAGENT SETUP

**Bacterial strains** *E. coli* BL21(λDE3). BL21 has the following genotype: $F^-$, *ompT*, *hsdS*$_B$ ($r_B^- m_B^-$) *gal*, *dcm*. Frozen competent cells of BL21(λDE3) can be obtained from many suppliers, including Stratagene, Promega and Genlanatis. BL21(λDE3) has a recombinant λ phage carrying the T7 RNA polymerase, stably integrated into the chromosome. ! **CAUTION** Use good microbiological practice in manipulating and disposing of this *E. coli* laboratory strain. BL21(λDE3) does not require additional amino acids to grow in minimal medium (MM). If you are using a different host strain, check the genotype and add amino acids as appropriate for growth. If you are unsure about the requirements, the supplier of the strain (or transformation-competent cells) should be able to provide you with a recipe for defined growth medium for a particular strain.

**Plasmids** The recombinant plasmids encoding the QconCAT proteins are called pQconCATs, and these carry QconCAT genes cloned into the *Nde*1-*Hin*dIII restriction sites of pET21a, although other restriction sites are available.

**Tris-EDTA buffer** Prepare 100 ml 10 mM Tris HCl, 1 mM EDTA, pH 8.0 (TE). Sterilize by autoclaving for 15 min at 121 °C, and store at room temperature (RT, 20–25 °C).

**LB** Dissolve 25 g of LB powder in 1 liter of distilled water. The pH should be 7.0 ± 0.2 at 25 °C; if it is not, adjust with HCl or NaOH as appropriate. Sterilize by autoclaving for 15 min at 121 °C.

**Ampicillin sodium salt** Prepare a 20 mg ml$^{-1}$ solution in sterile distilled water, freeze in 1-ml aliquots and store at –20 °C for several months, or store at 4 °C for no longer than 1 week. For the growth of BL21(λDE3)-pQconCAT strains use a final concentration of 50 μg ml$^{-1}$ to maintain selection for the recombinant plasmid.

**IPTG** Prepare a small volume of a 1 M solution in sterile distilled water.

**2× SDS-PAGE sample buffer** 0.5M Tris HCl, pH 6.8 (2.5 ml, for a final concentration of 0.125 M); 10% (wt/vol) SDS (4.0 ml, for a final concentration of 4% (wt/vol)); glycerol (2.0 ml (density 1.26 g ml$^{-1}$), for a final concentration of 20% (vol/vol)); DTT (0.31 g, for a final concentration of 0.2 M); bromophenol blue 0.5% (wt/vol), (40 μl). Add double-distilled water to the 2× SDS-PAGE sample buffer to a final volume of 10 ml, divide into 1-ml aliquots and store at –20 °C. ! **CAUTION** DTT solid is harmful by inhalation, in contact with skin and if swallowed. It is irritating to eyes, skin and the respiratory system.

**Tryptic digestion solution** 25 mM and 50 mM ammonium bicarbonate solution (no pH adjustment needed); Trypsin, sequencing grade. Prepare 10 ml of 10% (vol/vol) formic acid in distilled water for use in time course experiments. ! **CAUTION** Formic acid is harmful by inhalation; it causes severe burns.

**Solution for in-gel digestion and reduction and alkylation of cysteine-containing proteins** Prepare 20 ml of a 2:1 mixture of 25 mM ammonium bicarbonate–ACN.

**DTT** Prepare 10 ml of 10 mM DTT in 25 mM ammonium bicarbonate for in-gel digestion. Prepare 10 ml of 100 mM DTT stock solution in 25 mM ammonium bicarbonate for addition to analyte-QconCAT mixtures to a final concentration of 10 mM.

**Iodoacetamide** Prepare 10 ml of 55 mM iodoacetamide in 25 mM ammonium bicarbonate (cover the bottle with foil to exclude light) for in-gel digestion. ! **CAUTION** Iodoacetamide is toxic if swallowed. It may cause sensitivity by inhalation and skin contact. Do not breathe the dust. Wear suitable protective clothing and gloves.

**Matrix for matrix-assisted laser desorption-ionization–time of flight (MALDI-ToF) MS** Prepare 50 ml of 50% (vol/vol) acetonitrile–0.1% (vol/vol) TFA; store at RT. Prepare a fresh saturated solution of ~10 mg of α-cyano-4-hydroxycinnamic acid in 1 ml 50% (vol/vol) ACN–0.1% (vol/vol) TFA. ! **CAUTION** ACN is highly flammable and is harmful by inhalation, contact with skin and if swallowed. It is also irritating to the eyes. TFA is harmful by inhalation and causes severe burns. α-cyano-4-hydroxycinnamic acid is irritating to the eyes, respiratory system and skin.

**Solution for electrospray ionization (ESI) MS** Prepare 500 ml 50% (vol/vol) ACN–1% (vol/vol) formic acid. ! **CAUTION** Formic acid is harmful by inhalation; it causes severe burns.

**Guanidinium chloride binding buffer** 20 mM phosphate, pH 7.4; 0.5 M NaCl; 20 mM imidazole; 6 M guanidinium chloride.

**Elution buffer** 20 mM phosphate, pH 7.4; 0.5 M NaCl; 500 mM imidazole; 6 M guanidinium chloride.

**Dialysis buffer** 10mM ammonium bicarbonate, pH 8.5–1 mM DTT; the pH is adjusted with a solution of 7 M ammonium hydroxide.

**Protein assay** Coomassie Plus Protein Assay (Pierce); BSA for the protein calibration curve.

**Preparation of chicken skeletal muscle soluble fraction** Add one tablet Complete Protease Inhibitors to 20 ml of 20 mM sodium phosphate buffer, pH 7.0.

**Bacterial growth media** The minimal medium (MM) is a chemically defined mixture into which the stable isotope–labeled precursor to be used as a label can be added. For labeling with stable isotope–labeled amino acids, the following sterile stock solutions are made.

**5× M9 salts** To 90 ml of distilled water add the following, sequentially, allowing each salt to dissolve before adding the next: 6.4 g $Na_2HPO_4 \cdot 7H_2O$; 1.5 g $KH_2PO_4$; 0.25 g NaCl; 0.5 g $NH_4Cl$. Make up to 100 ml with distilled water, and autoclave at 121 °C for 15 min. Also prepare: 1 M $MgSO_4$ (autoclave); 20% (wt/vol) glucose (filter sterilize); 0.5% (wt/vol) thiamine (filter sterilize); 0.1 M $CaCl_2$ (autoclave).

**Amino acids** Prepare a 10 mg $ml^{-1}$ mixture including all the amino acids except those to be used for labeling. Vortex-mix the suspension vigorously (not all amino acids will dissolve fully), and immediately dispense in 2-ml aliquots and store at –20 °C (stable for several months). If you intend to use a variety of labeled amino acids in future experiments, omit all of these from this mixture so that you only have to make it once.

**MM with amino acids** Prepare MM by mixing the following volumes of the above stock solutions: 5× M9 salts (40 ml); 1 M $MgSO_4$ (0.2 ml); 20% (wt/vol) glucose (2 ml); 0.1 M $CaCl_2$ (0.2 ml); 0.5% (wt/vol) thiamine (20 µl); amino acid mix lacking chosen labeled amino acid(s) (2.0 ml); sterile water to make 200 ml.

**Labeled amino acid–supplemented MM** The medium used for growth should contain all 20 amino acids. The amino acid(s) chosen for labeling is weighed and added as a solid to the above MM plus amino acids. If you have left out other amino acids you should add these individually in unlabeled form.

Therefore, add 10 mg of either labeled or unlabeled amino acid per 100 ml of the MM with amino acids, filter sterilize and store at 4 °C for as long as 48 h.

**MM for labeling with $^{15}NH_4Cl$** For $^{15}NH_4Cl$ as the labeled precursor, it is essential that you choose an *E. coli* strain (like BL21(λDE3)) that does not require amino acid supplements to grow. Check the genotype of your host strain, and select a different host strain if appropriate. *E. coli* will grow more slowly in the absence of amino acids, so times required for growth and induction will be extended ~50%.

**$^{15}$N-labeled 5× M9 salts** $Na_2HPO_4$ (3.39 g; anhydrous $M_r$ 141.96); $KH_2PO_4$ (1.5 g); NaCl (0.25 g); $^{15}NH_4Cl$ (0.5 g); distilled water to make 100 ml. Autoclave in 20-ml aliquots.

**MM containing $^{15}$N** Mix the solutions as follows and store the medium at 4 °C for as long as 48 h: $^{15}$N-labeled 5× M9 salts (20 ml); 1 M $MgSO_4$ (0.1 ml); 20% (wt/vol) glucose (1 ml); 0.1 M $CaCl_2$ (0.1 ml). Add sterile distilled water to make 100 ml. Note: thiamine is not added, because this is a possible source of unlabeled nitrogen ($^{14}$N).

**EQUIPMENT**

- Standard spectrophotometer (for absorbance readings in the visible range, including 600 nm)
- Peristaltic pump
- MALDI-ToF MS instrumentation (e.g., Waters-Micromass Q-ToF micro mass spectrometer)
- ESI MS instrumentation
- HisTrap column (GE Healthcare)
- Multiskan plate reader (Thermo Electron)
- Ascent software
- Syringe pump (Harvard)
- MaxEnt 1 module of the MassLynx software
- Jouan centrifugal evaporator (Thermo Electron)

## PROCEDURE

### Transformation of expression host *E. coli* BL21(λDE3) with a QconCAT plasmid

**1|** Dissolve the QconCAT plasmid (pQconCAT), which is typically delivered as a lyophilized powder, in ~100 µl Tris EDTA buffer and store it at –20 °C.

**2|** Prepare a 1 ng/µl solution of the pQconCAT plasmid in Tris EDTA buffer.

**3|** Prepare BL21(λDE3) cells competent for transformation (using basic transformation methods; see ref. 19), or purchase frozen competent cells and introduce the pQconCAT plasmid by transformation using standard procedures (as described in the notes that accompany frozen competent cells).

**! CAUTION** Use good microbiological practice.

**4|** Select colonies on LA plates with 50 µg $ml^{-1}$ ampicillin by growing overnight at 37 °C.

**5|** Take a single colony and streak onto a fresh LA plate with 50 µg $ml^{-1}$ ampicillin. Store the freshly streaked plate at 4 °C for as long as a month.

### Preparation of frozen cultures of BL21(λDE3)-pQconCAT

**6|** Inoculate 10 ml of LB–50 µg $ml^{-1}$ ampicillin, with a single colony of BL21(λDE3)-pQconCAT, and grow overnight at 37 °C with shaking.

**7|** Add sterile glycerol to 30%, mix well, aliquot 1 ml into sterile microcentrifuge tubes and store at –70 °C.

■ **PAUSE POINT** Frozen cells may be kept in this way for several years.

**8|** To prepare a fresh plate, place an aliquot of frozen culture on ice and thaw a small amount of the top; use 20 µl to streak a fresh LA plate containing ampicillin, return the frozen culture to the freezer and incubate the plate overnight at 37 °C.

■ **PAUSE POINT** Plates can be kept for as long as a month under refrigeration, but it is best to inoculate cultures using a colony from a fresh plate.

### Induction of expression of QconCAT proteins in BL21(λDE3)-pQconCAT and confirmation of expression by SDS-PAGE

**9|** Using a single colony of BL21(λDE3)-pQconCAT, inoculate 10 ml LB containing ampicillin (50 µg $ml^{-1}$) and incubate overnight at 37 °C with shaking.

**10|** Transfer 500 μl of the overnight culture to 50 ml of prewarmed (to 37 °C) fresh LB medium (a 1:100 dilution) containing ampicillin (50 μg $ml^{-1}$), and incubate the culture with shaking. Remove 1-ml samples at hourly intervals, and determine the absorbance at 600 nm using a spectrophotometer.

**11|** When an $A_{600}$ of 0.6–0.8 is reached (usually ~2.5 h) add 50 μl of 1 M IPTG (final concentration of ~1 mM) to induce expression of the QconCAT protein.

**12|** Remove 1-ml samples at time 0 and then every 1–2 h (up to 6 h), measure the $A_{600}$ immediately, transfer the equivalent of 0.6–0.8 $A_{600}$ of cells to labeled microcentrifuge tubes and hold on ice.

**13|** Centrifuge the samples at 8,000*g* in a microfuge at 4 °C for 10 min. Remove and discard all the supernatant using a micropipette, and suspend the pellet in 100 μl of distilled $H_2O$ by vigorous vortexing.
■ **PAUSE POINT** If the intention is to perform the SDS-PAGE analysis at a future date, place at −20 °C at this stage.

**14|** Transfer the remainder of the culture to a preweighed 50-ml centrifuge tube, and centrifuge at 1,450 *g* for 10 min at 4 °C.

**15|** Decant the supernatant, and weigh the tube again to determine the wet weight of the cell pellet. Freeze at −20 °C until required for purification of QconCAT after confirmation by SDS-PAGE of sufficient levels of induction.

**16|** Add 100 μl double-strength SDS sample buffer to samples from Step 13, mix well and place in a boiling water bath for 4 min. Analyze 20 μl of each sample by SDS-PAGE mini-gel, carry out electrophoresis, and stain and destain the gel (e.g., with Coomassie blue).

**17|** The QconCAT protein should appear as a clearly visible band that is either not present or much fainter in the time 0 sample. Check that the molecular weight is close to that calculated for the QconCAT, and if all is as expected proceed to the next stage.

**Analysis of QconCAT by MALDI-ToF MS after in-gel digestion of SDS-PAGE gel band with trypsin**

**18|** Using a Pasteur pipette, cut a plug of gel from the band corresponding to the QconCAT protein, transfer the plug to a 1.5-ml microcentrifuge tube, add 25 μl of 25 mM ammonium bicarbonate and incubate at 37 °C for 15 min. Discard any liquid.

**19|** Add 25 μl of 25 mM ammonium bicarbonate–ACN (2:1), and incubate at 37 °C for 15 min. Discard any liquid.
! **CAUTION** ACN is flammable and toxic.

**20|** Add 25 μl of 25 mM ammonium bicarbonate, and incubate at 37 °C for 15 min. Discard any liquid. Repeat Steps 19 and 20 using alternate washes until the plug is fully destained. Note: If cysteine residues are present, alkylate the QconCAT protein by performing these three steps: (i) Add 25 μl of 10 mM DTT in 25 mM ammonium bicarbonate, and incubate at 56 °C for 60 min. Discard any liquid. (ii) Add 25 μl 55 mM iodoacetamide in 25 mM ammonium bicarbonate, and incubate in the dark at 37 °C for 45 min. Discard any liquid. (iii) Repeat alternate washes (Steps 19 and 20), finishing with a wash in 25 mM ammonium bicarbonate–ACN (2:1). These three steps can be omitted if cysteine residues are not present in the QconCAT protein.

**21|** Add 10 μl of 25 mM ammonium bicarbonate containing 0.1 mg $ml^{-1}$ trypsin giving a final concentration of 12.5 ng $\mu l^{-1}$.

**22|** Hold samples on ice for 30 min before incubation at 37 °C overnight. Ensure that the gel plug is covered with liquid; if it is not, add 25 mM ammonium bicarbonate until covered.

**23|** Spot 1 μl of digest on a MALDI target, and overlay with 1 μl of matrix.

**24|** Acquire spectra using a MALDI-ToF mass spectrometer. Acquire mass spectra over the range 800—3,500 *m/z*.

**Evaluation of MALDI ToF results**

**25|** Analyze spectra. Peaks due to each peptide in the QconCAT with masses larger than 800 *m/z* should be clearly visible. Confirm that all the expected peaks are present, and confirm that digestion has gone to completion—meaning there are no peaks due to missed cleavages.

**Purification and analysis of QconCAT proteins**

**26|** Thaw the induced cells from Step 15. The QconCAT is generally present in inclusion bodies, but it is important to confirm this. Inclusion bodies are first recovered by breaking cells using BugBuster Protein Extraction Reagent (Novagen). The following method is adapted from the Novagen protocol.

**27|** For ≤1 g of wet cell pellet add 2.5 ml of BugBuster (BB) and, to ensure good resuspension, place cells on a rocker platform at RT for 15 min. Remove 20 μl for analysis by SDS-PAGE (total fraction).

**28|** Centrifuge the cells at 16,000*g* for 20 min at (4 °C). Set the braking speed low to give a gentle rotor deceleration.

**29|** Collect the supernatant by pipetting carefully to a fresh tube, transfer 20 µl to a labeled microcentrifuge tube for analysis (soluble fraction) and store the remainder at −20 °C. Resuspend the pellet in 2.5 ml of BB by drawing the cells up into a Pasteur pipette and by gentle vortexing.
▲ **CRITICAL STEP** Thorough resuspension is critical to obtaining a preparation of high purity.

**30|** Add 50 µl of lysozyme (10 mg $ml^{-1}$ in BB), mix gently by vortexing and incubate at RT for 5 min.

**31|** Add 15 ml of 1:10 dilution of BB in distilled water, and mix by vortexing for 1 min.

**32|** Centrifuge at 15,000*g* for 15 min at 4 °C.
▲ **CRITICAL STEP** This is a much higher speed than recommended in the Novagen protocol, but it generates a firmer pellet, which facilitates handling. Carefully decant the supernatant and discard.

**33|** Resuspend the pellet of inclusion bodies in 20 ml of 1:10 BB, and mix by vortexing (low speed), then centrifuge at 15,000*g* for 15 min at 4 °C. Discard the supernatant.

**34|** Repeat Step 33 twice more, finishing by centrifuging at 16,000*g* for 15 min. Store the pellet of inclusion bodies at −20 °C in the 50-ml centrifuge tubes.
■ **PAUSE POINT** Inclusion body pellets can be stored at −20 °C indefinitely.

**35|** Analyze the total fraction (Step 27) and soluble fraction (Step 29) by SDS-PAGE.

**36|** Add 20 µl of 2× SDS-PAGE sample buffer to each sample, heat in a boiling-water bath for 4 min and load 20 µl onto a gel.

**37|** From the Coomassie blue–stained gel determine the location of the QconCAT and, if largely absent from the soluble fraction (compared with the total fraction), proceed with the purification of the QconCAT protein from the inclusion bodies.

**38|** Resuspend the inclusion bodies in 20 ml of binding buffer at room temperature. (If the QconCAT is in the soluble fraction, add this fraction to 20 ml of binding buffer.)

**39|** Centrifuge at 8,000 r.p.m. (~5,000*g*) for 5 min at room temperature, and collect the supernatant in a fresh tube. Place 20 µl of the supernatant into a microcentrifuge tube; this sample is the starting material (SM) for SDS-PAGE analysis.

**40|** Apply the remainder of the supernatant, with the aid of a peristaltic pump set at a flow rate of 0.25 ml $min^{-1}$, to a 1-ml HisTrap HP column equilibrated in the same buffer.

**41|** Collect the unbound sample in a 25-ml Universal tube. Place 200 µl of the eluate into a microcentrifuge tube and hold at room temperature; this sample is the unbound material (UM).

**42|** Wash the column with 10 ml of 20 mM phosphate, pH 7.4, 20 mM imidazole, 0.5 M NaCl, 6 M guanidinium chloride using an increased flow rate of 0.5 ml $min^{-1}$, and collect the wash. Transfer 200 µl of the wash into a microcentrifuge tube, and hold at room temperature; this sample is the wash (W).

**43|** Elute the bound protein with 5 ml 20 mM phosphate, pH 7.4, 500 mM imidazole, 0.5 M NaCl, 6 M guanidinium chloride at a flow rate of 0.25 ml $min^{-1}$, and collect five 1-ml fractions. Place 20 µl of each fraction into a microcentrifuge tube, and hold at room temperature; this sample is the eluted bound material (EBM). Hold the eluted fractions on ice.
■ **PAUSE POINT** Samples can be stored at −20 °C.

**44|** To identify the fractions containing the QconCAT, prepare each of the 20-µl fractions for SDS-PAGE analysis. Collect the samples SM, UM, W and EBM (Steps 39–43), and add 10 µl of a StrataClean Resin (Stratagene) bead suspension to each tube (use the same volume of beads for the 200-µl W and UM samples).
▲ **CRITICAL STEP** The guanidinium chloride interferes with the SDS-PAGE and so must be removed.

**45|** Vortex each sample for 1 min, centrifuge for 2 min at 230 *g* at RT, and remove and discard the supernatant.

**46|** Resuspend pellets in 1 ml of distilled $H_2O$, vortex briefly to mix and centrifuge at 230*g* for 2 min. Discard the supernatant.

**47|** Add 10 µl of 2× SDS sample buffer to the beads, boil for 4 min and then load both sample and beads onto the gel.

**48|** Stain (with Coomassie blue) and destain the gel.

**49|** Pool the eluant fractions containing pure QconCAT protein, and remove denaturant by dialyzing against 100 volumes of 10 mM ammonium bicarbonate, pH 8.5, 1 mM DTT, at 4 °C, for 2 h. Repeat twice more with fresh buffer.

**50|** Determine the protein concentration. For this one can use Coomassie Plus Protein Assay (Pierce) and BSA as the standard, and a LabSystems Multiskan plate reader using Ascent software.

**51|** Prepare 1:50 and 1:100 dilutions of the purified QconCAT protein with a protein assay ranging from 0 to 50 μg. The concentration of the QconCAT protein is likely to be in the range of 0.2–4 mg $ml^{-1}$.

**Analysis of QconCAT proteins using ESI MS (Q-ToF)**

**52|** Dilute purified and dialyzed QconCAT protein directly into 50% (vol/vol) ACN–1% (vol/vol) formic acid to a final concentration of 60–100 fmol $\mu l^{-1}$.

**53|** Directly infuse the sample into the source at a flow rate of 0.5 μl $min^{-1}$ using a syringe pump (Harvard). The capillary voltage may be set between 1,600 and 2,100 V and data acquired over a mass range of 400–1,500 *m/z* with a scan/interscan speed of 2.4/0.1 s.

**54|** Combine the scans and subtract the spectra before transformation using the MaxEnt 1 module of the MassLynx software. This permits deconvolution of the spectrum into an intact mass measurement within 2 Da of the predicted mass.

**Growth of BL21(λDE3)-pQconCAT in minimal medium and labeling with stable isotopes**

**55|** Using sterile technique, pipette 10 ml of MM without amino acids into a sterile 50-ml conical flask, and add 25 μl of 20 mg $ml^{-1}$ ampicillin in sterile distilled water.

**56|** Inoculate with a single colony of BL21(λDE3)-pQconCAT, and incubate overnight at 37 °C with shaking.

**57|** Using sterile technique, transfer 50 ml MM plus amino acids and 50 ml MM plus labeled amino acids to two sterile 250-ml conical flasks. Add 125 μl of 20 mg $ml^{-1}$ ampicillin in sterile distilled water to each, and warm to 37 °C.
▲ **CRITICAL STEP** Amino acids are included in this step to increase the growth rate in MM. This shortens the time before induction can commence and prevents the recycling of labeled amino acids.

**58|** Determine the $A_{600}$ of the overnight culture and, using sterile technique, transfer enough culture to each flask to give a starting $A_{600}$ of 0.06–0.1.

**59|** Monitor the $A_{600}$ of the cultures at hourly intervals until an $A_{600}$ of 0.6–0.8 is reached (this should take 3.5–4.5 h depending on the starting $A_{600}$, the *E. coli* strain and the growth medium used).

**60|** Add 50 μL 1 M IPTG to each flask to induce protein expression. Remove a 1-ml sample at 0 and 5 h after induction, measure the $A_{600}$ and prepare for gel analysis exactly as described in Steps 13, 16 and 17.

**61|** Centrifuge the remainder of the 5-h induced cells in preweighed centrifuge tubes, at 1,450 *g* for 15 min at 4 °C, decant and discard the supernatant, record the weight of the cell pellets and store the cell pellets in the centrifuge tubes at −20 °C.

**62|** Process exactly as described in Steps 26–54, for purification and analysis.

**Comparison of labeled and unlabeled QconCAT tryptic peptide profiles**

**63|** Mix unlabeled (L, light) and labeled (H, heavy) QconCAT protein to give a total of 10 μg (e.g., L/H ratios of 10:0, 9:1, 7:3, 5:5, 3:7, 1:9 and 0:10), and add 50 mM ammonium bicarbonate to a volume of 20 μl.

**64|** Add trypsin in a ratio of trypsin to QconCAT of 1:50, and incubate at 37 °C overnight.

**65|** Analyze four 1-μl fractions by MALDI-ToF MS as described in Steps 23 and 24.

**66|** Measure the peak intensities, and confirm that the quantification is as expected from the prepared ratios.

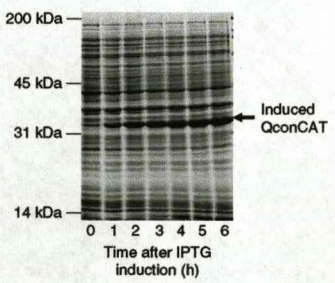



**Figure 4 |** Induction of QconCAT protein in BL21(λDE3) cells grown in Luria broth. BL21(λDE3)-pQconCAT cells were grown to an $A_{600}$ of 0.8, and IPTG added to a final concentration of 1 mM. Cells equivalent to 0.08 $A_{600}$ units were analyzed by SDS-PAGE and Coomassie blue staining.

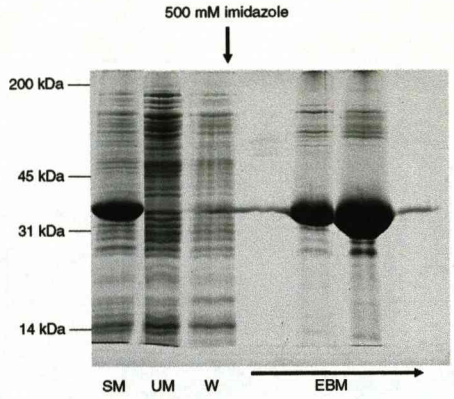



**Figure 5** | HisTrap purification of QconCAT. SM, starting material (solubilized inclusion bodies), UB, unbound material (flowthrough); W, wash fraction; EBM, eluted bound material (sequential 1-ml fractions). The middle two fractions containing the eluted QconCAT protein are pooled for use.

**Determination of conditions required for complete digestion of the intended analyte and confirming the digestion of analyte and QconCAT protein mixtures**

**67|** Label 12 0.5-ml microcentrifuge tubes, pipette 10 µl of 10% (vol/vol) formic acid into each tube and hold on ice.

**68|** Dilute ~20 µg of analyte protein 1:10 with 50 mM ammonium bicarbonate solution. This typically gives a volume of 200–300 µl.

**69|** Add trypsin in a ratio of trypsin to analyte of 1:50, mix and immediately transfer 12 µl into the first tube previously prepared in Step 67.

**70|** Incubate the analyte-trypsin mixture at 37 °C, and take 12-µl samples over a 24-h period. Suitable time intervals are 0, 1, 2, 5, 10, 30, 60, 90, 120, 240, 480 and 1,440 min. Store samples at –20 °C.

**71|** Thaw the samples, and spot 1 µl of each sample onto two positions on a MALDI target; that is, perform in duplicate. Analyze by MALDI-ToF MS exactly as described in Steps 23 and 24. If digestion is complete, the peptides of the proteins under study (i.e., chosen for inclusion in the QconCAT) should be readily seen. At early time points partial digestion products will be seen for some proteins; this analysis allows the monitoring of these more slowly digested portions of the protein, which should be fully digested within 24 h if conditions are appropriately optimized.

**72|** Dry down the remaining sample (10 µl) for 2 h by heating under vacuum (Jouan centrifugal evaporator (Thermo Electron)), and resuspend the residue directly in 10 µl 1× SDS sample buffer.

**73|** Heat samples in a boiling-water bath for 4 min, and load the entire sample onto an 12.5 % SDS-PAGE gel. After electrophoresis stain the gel with Coomassie blue; no protein should be visible on the gel after 24 h of digestion.

**74|** If digestion of the analyte is incomplete, increase the protein-to-trypsin ratio from 50:1 to 20:1; also, an organic solvent such as ACN can be added (e.g., to 10%). Complete digestion of the analyte proteins is measured by the absence of all bands in an SDS-PAGE gel.

**75|** Once you have optimized digestion conditions for the analyte, repeat the digestion but add the labeled QconCAT protein in a 1:10 ratio with the analyte.

**76|** Analyze 1 µl by MALDI-ToF MS, and confirm the complete digestion of the QconCAT when present in a mixture with the analyte.

**Quantification of QconCAT proteins for use in absolute quantification**

**77|** Quantify the labeled QconCAT by a method of suitable accuracy for the intended experiment (see section on "Quantification of the QconCAT" in the EXPERIMENTAL DESIGN section). So far we have used a dye-binding assay to determine the protein concentration of both the QconCAT and the analyte proteins. This method gives acceptable accuracy and does not require the use of significant quantities of the

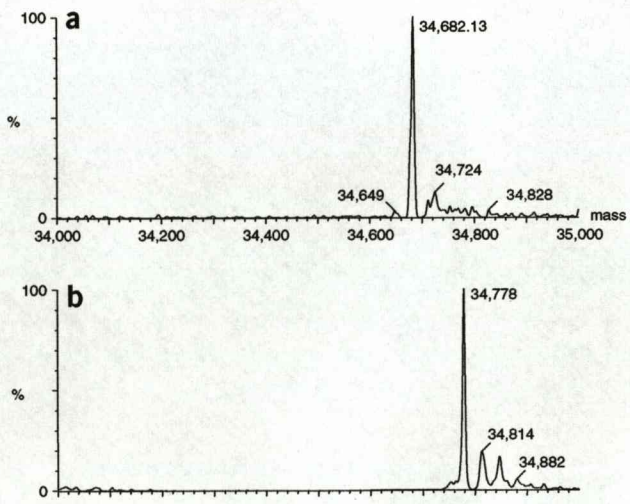



**Figure 6** | ESI-MS to measure the intact mass of QconCATs. Purified and dialyzed, labeled and unlabeled, QconCAT protein was diluted into 50% (vol/vol) ACN–1% (vol/vol) formic acid to a final concentration of 60–100 fmol $\mu l^{-1}$ and infused directly into the source of a Q-ToF MS at a flow rate of 0.5 µl $min^{-1}$ using a syringe pump (Harvard). The capillary voltage was set between 1,600 and 2,100 V and data acquired over a mass range of 400–1,500 *m/z* with a scan/interscan speed of 2.4/0.1 s. Scans were combined and the spectra subtracted before transformation using the MaxEnt 1 module of the MassLynx software, which allows deconvolution of the spectrum into an intact mass measurement within 2 Da of the predicted mass. (**a**) Unlabeled QconCAT (predicted mass: 34,684); (**b**) QconCAT labeled with [$^{13}C_6$]lysine (predicted mass 34,780).

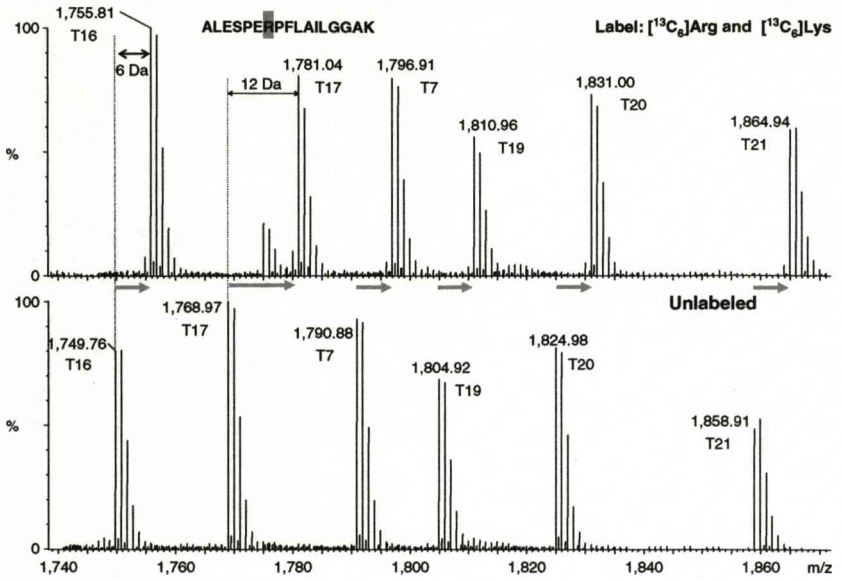



**Figure 7** | Incorporation of label into peptides derived from QconCATs. QconCAT proteins labeled with [$^{13}C_6$ ]lysine and [$^{13}C_6$ ]arginine (upper spectrum) and unlabeled (lower spectrum) were digested with trypsin and each digestion analyzed directly by MALDI-ToF MS. Only a portion of the spectrum is shown so that the mass shift of the labeled peptides can be seen. Peptide T17 differs from its unlabeled counterpart by 12 Da as a result of the uncleaved arginine-proline sequence in the peptide; all the other peptides show a shift to 6 Da heavier in the labeled peptide.

purified QconCAT. We are currently exploring the use of other methods that are more accurate but utilize small amounts of the purified QconCAT.

### Use of a QconCAT protein in absolute quantification: quantification of proteins in chicken skeletal muscle

**78|** Take 100 mg of chicken breast tissue from chickens of different ages (for example), and homogenize in 1 ml 20 mM sodium phosphate buffer, pH 7.0, containing protease inhibitors.

**79|** Centrifuge at 15,000*g* for 45 min at 4 °C. Collect the soluble fraction, and store at −20 °C in 100-µl aliquots. It may be necessary to rehomogenize/recentrifuge the initial pellet to recover all of the soluble material.

**80|** Determine the total protein concentration in each sample of chicken skeletal muscle soluble fraction (CSM) and the concentration of the labeled purified QconCAT protein using a Coomassie Plus Protein Assay.

**81|** Dilute samples 1:10 with 50 mM ammonium bicarbonate, and mix the CSM and labeled QconCAT in known concentrations. For initial pilot studies a ratio of 1:10 (QconCAT/CSM) is suitable, and total protein in the range of 20 µg.

**82|** Add ACN to 10% and trypsin in a 20:1 (protein to enzyme) ratio, and incubate at 37 °C for 24 h.

**83|** Spot four 1-µl (four replicates) of each digest onto a MALDI target, and overlay with 1 µl of matrix.

**84|** Acquire spectra, for example using a MALDI-ToF mass spectrometer, over the range 800—3,500 *m/z*, or use a lower *m/z* maximum if peptides under study fall well below this maximum mass.

**85|** Record the intensity of the monoisotopic peak for each analyte peptide and the corresponding QconCAT peptide, and convert these values into nanomoles of protein per gram of tissue using the known concentration of the QconCAT protein. Determine the standard error for each ratio.

### ? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

**TABLE 1** | Troubleshooting the QconCAT design, expression and labeling steps.

| PROBLEM | POSSIBLE REASON | SOLUTION |
|---|---|---|
| Expression of the QconCAT is very poor | Multiple potential reasons, see below | If expression is low but just detectable, sufficient QconCAT protein can be obtained by processing larger volumes of induced cultures. |
| | Gene sequence incorrect | Check the gene sequence. Look at transcript and predicted translation product. If incorrect, the gene must be reconstructed. |
| | Protein instability | Vary the induction conditions (times, temperature) to try to improve expression. Try a different host *E. coli* strain. |
| | Problems with mRNA secondary structure occluding expression | Monitor RNA production upon induction to confirm that transcription is proceeding well. If this is not the problem, alter the order of the peptides near the N terminus of the QconCAT to generate a new gene construct. |

**TABLE 1** | Troubleshooting table (continued).

| | | |
|---|---|---|
| | Protein toxicity | If the expression problem is due to toxicity the cells are likely to lyse upon induction; monitor the $A_{600}$ of the culture during the induction phase. Alter the order of the peptides near the N terminus of the QconCAT to generate a new gene construct. |
| Poor ion intensities of selected Q-peptides | Many of the peptides terminate with lysine | Specific to MALDI-ToF MS. Investigate whether guanidination of the QconCAT peptides overcomes the problem[10–12]. |
| | Peptide ionizes poorly by chosen MS ionization method | Try alternative ionization methods. Peptides that are known to yield poor signals in MALDI-ToF may perform well in ESI MS. |
| | Peptide chosen ionizes very weakly using any MS approach | Select another peptide to represent the protein in question. This requires generating a new gene. Fortunately this only involves changing a few of the oligonucleotides used for gene synthesis, and a new version can be rapidly and economically generated. |
| Incomplete labeling of QconCAT protein with stable isotopes | Precursor pool is incompletely labeled. Either the isotope has a relative isotope abundance less than 0.99 or endogenous, unlabeled amino acids are diluting the labeled precursor pool. | |
| | Minimal medium is contaminated with unlabeled versions of heavy-isotope precursors. | Re-prepare minimal medium from pure reagents. |
| | Too large an inoculum of the overnight culture grown in MM plus all amino acids was used to start the culture for the labeling step. | Decrease the size of the inoculum. |
| | Insufficient heavy isotope–labeled precursor is added, forcing *E. coli* to synthesize amino acids *de novo*, which are not labeled. | Add more heavy-isotope precursor for growth and induction, or use an auxotroph for the amino acid. |
| Incomplete digestion of analyte proteins and QconCAT | Insufficient trypsin | Increase ratio of trypsin to analyte proteins. Alter digestion conditions to enhance proteolysis. Pretreat analyte proteins to increase susceptibility to digestion. |
| | Conditions suboptimal | Add organic solvent (e.g., ACN) to 10%. |
| | Proteins are tightly folded | Include a denaturant (e.g., urea 2–4 M), and increase the trypsin concentration because the denaturant will diminish trypsin activity. |
| | Proteins are very hydrophobic membrane proteins, with cleavage sites unavailable due to steric hindrance or sequestration within membrane vesicles | Solubilize membrane preparations using, for example, organic solvents, organic acids or detergents[7–9] (e.g., nonionic detergent such as Rapigest; Waters (http://www.waters.com)). |



## ANTICIPATED RESULTS

During the course of the induction of expression of QconCAT proteins, a band corresponding to the expected molecular weight of the QconCAT protein should be readily seen to increase in intensity, confirming good levels of expression (**Fig. 4**).

Purification and SDS-PAGE analysis confirms that the QconCAT was readily solubilized from the inclusion bodies, bound efficiently to the HisTrap column (i.e., was not present in the unbound fraction) and was the major band in the bound eluted fractions (**Fig. 5**).

ESI MS analysis of the intact QconCAT protein allows comparison of the experimentally derived molecular weight and the calculated molecular weight (**Fig. 6a**). Loss of the N-terminal methionine can be predicted, and the degree to which other potential modifications to the primary structure have occurred can be evaluated (e.g., oxidation of methionine). Analysis of the stable isotope–labeled QconCAT protein allows evaluation of the labelling step (**Fig. 6b**). Protein peaks that are present between the unlabeled and labeled protein predicted masses could indicate incomplete labeling with heavy isotope precursors.

MALDI-ToF analysis of labeled and unlabeled QconCATs (in this case labeled with [$^{13}C_6$ ]lysine and [$^{13}C_6$]arginine) allows confirmation of the consistent 6-Da difference in mass between labeled and unlabeled forms of each peptide, unless an uncleavable sequence (e.g., RP) is present, as is the case with peptide T17, which differs from its unlabeled counterpart by 12 Da (**Fig. 7**).

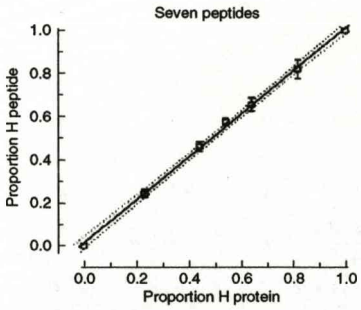



**Figure 8** | Quantification by QconCATs. Unlabeled (L, light) and uniformly labeled with $^{15}N$ (H, heavy) QconCAT proteins were separately purified, quantified and mixed in different ratios, before tryptic digestion and measurement of peptide intensities by MALDI-ToF MS. The measured proportion of H peptide is plotted relative to the proportion of H protein in the mixture for three replicates of each of seven peptides; error bars ± s.e.m., $n = 7$. The dotted lines define the 95% confidence limits of the fitted line. (Adapted from reference 6.)

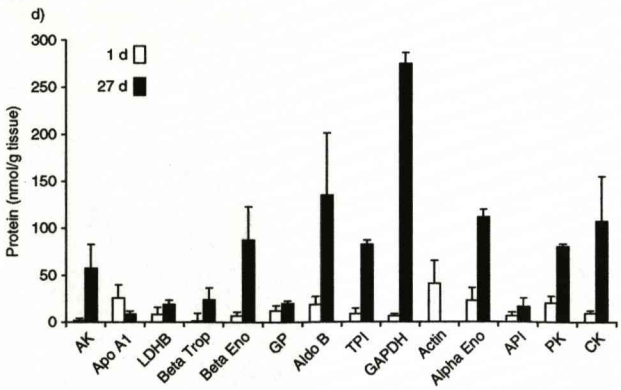



**Figure 9** | Quantification of biological samples by QconCAT. A preparation of soluble proteins from 10 mg of chicken skeletal muscle at 1 d and 27 d was mixed with 29 μg $^{15}N$-labeled QconCAT, digested with trypsin overnight and analyzed by MALDI-ToF. The intensity of the monoisotopic peak for the analyte peptide and the corresponding QconCAT peptide were recorded and the data converted into nanomoles of protein per gram of tissue, to give the absolute amount of each protein. Error bars, ± s.e.m.; $n = 3$. For details of the individual proteins, see reference 6. (Adapted from reference 6.)

Statistical analysis of labeled and unlabeled QconCAT proteins, mixed in known proportions, digested and analyzed by MALDI-ToF MS (**Fig. 8**) allows evaluation of the reproducibility of the MALDI-ToF analysis. In addition, the behavior of each peptide within the QconCAT can be assessed and the 95% confidence limits, of the line fitted to the data and visualized.

The data shown in **Figure 9** are adapted from ref. 6 and show the absolute quantification values, as well as standard errors, that can be obtained using QconCAT proteins as internal standards. Results are presented as nanomoles of each protein per gram of tissue.

**ACKNOWLEDGMENTS** This work was supported by the Biotechnology and Biological Sciences Research Council, Swindon, UK.

**COMPETING INTERESTS STATEMENT** The authors declare competing financial interests (see the HTML version of this article for details).

Published online at http://www.natureprotocols.com/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Julka, S. & Regnier, F.E. Recent advancements in differential proteomics based on stable isotope coding. *Brief. Funct. Genomic. Proteomic.* **4**, 158–177 (2005).
2. Ong, S.E. & Mann, M. Mass spectrometry–based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
3. Yan, W. & Chen, S.S. Mass spectrometry–based quantitative proteomic profiling. *Brief. Funct. Genomic. Proteomic.* **4**, 27–38 (2005).
4. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* **100**, 6940–6945 (2003).
5. Kirkpatrick, D.S., Gerber, S.A. & Gygi, S.P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* **35**, 265–273 (2005).
6. Beynon, R.J., Doherty, M.K., Pratt, J.M. & Gaskell, S.J. Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat. Methods* **2**, 587–589 (2005).
7. Wu, C.C. & Yates, J.R. The application of mass spectrometry to membrane proteomics. *Nat. Biotechnol.* **21**, 262–267 (2003).
8. Lu, X. & Zhu, H. A novel proteomic approach for high throughput analysis of membrane proteins. *Mol. Cell Proteomics* **4**, 1948–1958 (2005).
9. Fischer, F. & Poetsch, A. Protein cleavage strategies for an improved analysis of the membrane proteome. *Proteome Sci.* **4**, 1–12 (2006).
10. Brancia, F.L. *et al.* A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis* **22**, 552–559 (2001).
11. Beardsley, R.L. & Reilly, J.P. Optimization of guanidination procedures for MALDI mass mapping. *Anal. Chem.* **74**, 1884–1890 (2002).
12. Thevis, M., Ogorzalek Loo, R.R. & Loo, J.A. In-gel derivatization of proteins for cysteine-specific cleavages and their analysis by mass spectrometry. *J. Proteome Res.* **2**, 163–172 (2003).
13. Beynon, R.J. & Pratt, J.M. Metabolic labelling of proteins for proteomics. *Mol. Cell Proteomics* **4**, 857–8872 (2005).
14. Ellison, D., Hinton, J., Hubbard, S.J. & Beynon, R.J. Limited proteolysis of native proteins: the interaction between avidin and proteinase K. *Protein Sci.* **4**, 1337–1345 (1995).
15. Hubbard, S.J. The structural aspects of limited proteolysis of native proteins. *Biochim. Biophys. Acta* **1382**, 191–206 (1998).
16. Hubbard, S.J., Beynon, R.J. & Thornton, J.M. Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.* **11**, 349–359 (1998).
17. Wu, C., Robertson, D.H., Hubbard, S.J., Gaskell, S.J. & Beynon, R.J. Proteolysis of native proteins. Trapping of a reaction intermediate. *J. Biol. Chem.* **274**, 1108–1115 (1999).
18. Hubbard, S.J. & Beynon, R.J. Proteolysis of native proteins as a structural probe. In *Proteolytic Enzymes. A Practical Approach* (eds. Beynon, R.J. & Bond, J.S.) 233–264 (Oxford University Press, Oxford, 2001).
19. Sambrook, J. & Russell, D.W. (eds.) Preparation and transformation of competent *E. coli* using calcium chloride (Protocol 25). In *Molecular Cloning: A Laboratory Manual* 3rd Edn. Vol 1. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001).