Proteomic Applications of Protein and Peptide Transamination

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

By

Hao Zhang

April 2019

Abstract

The extreme amino (N)-terminus of a protein is highly informative of protein stability, localisation, function, as well as regulation. Protein N-termini often undergo post-translational modifications (PTMs), some of which generate new proteoforms with *neo*-N-termini due to proteolytic processing. N-terminal PTMs and the resulting proteoforms greatly contribute to the complexity of the proteome. Additionally, authentic N-termini may differ from those predicted by genome analysis.

This thesis describes studies to further develop and to optimise techniques that modify, enrich and identify the N-termini of proteins. The previously reported "NHS-Sepharose technique" removes unwanted peptides after acetylation of amino groups and protease digestion, thereby enriching for N-terminal peptides. Using this negative selection strategy, the desired N-terminal peptides were shown to be contaminated with specific *neo*-peptides. Various parameters were therefore systematically modified in conjunction with model proteins to achieve complete removal of *neo*-peptides. Investigation of alternative amine-reactive chemistries using *N*-succinimidyl *S*-acetylthioacetate or citraconic anhydride indicated that they were not superior to the original amine acetylation approach.

The studies also explored a positive selection strategy that directly enriches for protein N-termini. Selective Transamination Of N-Ends (STONE) is a promising reaction for this purpose. Use of STONE with subsequent biotinylation of the introduced carbonyl groups allowed N-terminal affinity tagging of synthetic peptides and model proteins. It was confirmed that the latter were amenable to avidin-biotin affinity enrichment. Treatment of extracts from Jurkat T-lymphocytes by STONE allowed the experimental verification of N-termini from a range of proteins. Previously unreported N-terminal methionine excision was also uncovered in cytoplasmic serine-tRNA ligase (SYSC_HUMAN) and peptidyl-prolyl cis-trans isomerase FKBP5 (FKBP5_HUMAN).

STONE-mediated removal of the α-amino groups of peptides and proteins alters their net charge. This suggested that it could be used to achieve orthogonal peptide separations in LC-MS/MS, and hence discover additional peptides in shotgun proteomic experiments. Using protein extracts from Jurkat T-cells, STONE was shown to increase the number of identified peptides by 25%, thereby significantly also increasing protein sequence coverage and the number of assigned proteins. The unexpectedly large increase in the number of detected peptides, coupled with the versatility and simplicity of STONE, suggests that it may find widespread use in proteomics.

Acknowledgements

I would like to thank my supervisor, Prof. David O'Connor for leading me into the fascinating world of science. I have been dreaming of working in proteomics since the third year of my undergraduate study, so this is indeed a dream come true. I am sincerely grateful to his supervision and guidance during the past five years. Most importantly, I thank him for the support when I met so many unexpected encounters in my life. All in all, it has been a fulfilling experience to work in his lab.

I would also like to thank my second supervisor, Prof. Rob Beynon, and two PhD assessors, Prof. Claire Eyers and Dr David Ruiz-Carrillo, for the inspirational discussions about my research. I also need to acknowledge both Xi'an Jiaotong-Liverpool University and the University of Liverpool for providing this opportunity to do my research and the financial support.

I would like to express my sincere gratitude to the past and present members in the DO'C lab: Jing Li, Yanni Xue, Zimeng Zhang, Li Shen, Yu Gan, and Manting Xue. In particular, I truly thank Jing Li for sharing the lab knowledge and helping me analyse all the data. I would like to acknowledge Zimeng Zhang and Yu Gan for the participation in systematic refinement of the NHS-Sepharose method and investigation of alternative amine-reactive chemistries, respectively. All the data were jointly analysed. I am also grateful to Li Shen for our mutual understanding and the crazy talks about science.

In addition, I am truly grateful to other research groups in the Department of Biological Sciences at XJTLU. In particular, I thank Prof. Tatsuhiko Kadowaki and Dr Magdalini Matziari for providing reagents and valuable suggestions on my research. I appreciate Dr Hebin Liu and Dr Rong Rong for providing Jurkat cell line.

I am sincerely grateful to the PhD community in the Department, including Dr Xiaochen (Will) Liu, Dr Xiaofeng Dong, Dr Fan Bu, Dr Adharsh Rajasekar, Dr Kiran Kumar, and Dr Dongqing Ma. In particular, I am grateful to Zhen Wei and Jingting Zhu for the training on statistical analysis with R and cell culture skills, respectively.

I cannot give enough credits to my friends, Guangcai Xu, Cheng Jiang, Bingjie Jiang, Liji Huang, who were always turning my complaints into positive thinking. Finally, my eternal gratitude goes to my parents and my wife for their unconditional love. My biggest regret during the past five years is not being with my mother when she suffered so much for me.

Table of Contents

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
2PCA	2-Pyridinecarboxaldehyde
ACTH	Adrenocorticotropic hormone
AD	Alzheimer's disease
AI	Artificial intelligence
ANOVA	Analysis of variance
AP	Affinity purification
APC	Antigen-presenting cell
APEX	Absolute protein expression
AP-MS	Affinity purification–mass spectrometry
AQUA	Absolute quantification
BCA	Bicinchoninic acid
Bis-Tris	Bis(2-hydroxyethyl)amino-tris(hydroxymethyl)methane
$BnONH_2$	*O*-Benzylhydroxylamine
bp	Base-pair
BSA	Bovine serum albumin
CA	Citraconic anhydride
CD	Cluster of differentiation
ChaFRADIC	Charge-based fractional diagonal chromatography
CID	Collision-induced Dissociation
COFRADIC	Combined fractional diagonal chromatography
DAVID	Database for Annotation, Visualization and Integrated Discovery
DC	Direct current
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNPH	2,4-Dinitrophenylhydrazine
DTT	DL-Dithiothreitol
***E*-value**	Peptide expectation value
EDTA	Ethylenediaminetetraacetic acid
emPAI	Exponentially modified protein abundance index
ESI	Electrospray ionisation
ETD	Electron Transfer Dissociation

FBS	Foetal bovine serum
FDR	False discovery rate
GO	Gene Ontology
h	Hour
HCD	Higher-energy Collisional Dissociation
HD	Huntington's disease
HPG-ALD	Hyperbranched polyglycerol-aldehydes
iBAQ	Intensity-based absolute quantification
ID	Identifier (Swiss-Prot or ProteomeXchange)
IEF	Isoelectric focusing
IgG	Immunoglobulin G
IL-2	Interleukin-2
iTRAQ	Isobaric peptide tags for relative and absolute quantification
K_d	Dissociation constant
LC	Liquid chromatography
LC-MS/MS	Liquid chromatography–tandem mass spectrometry
m	Mass
MALDI	Matrix-assisted laser desorption/ionisation
MES	2-(*N*-Morpholino)ethanesulfonic acid
MetAP	Methionine aminopeptidase
MGF	Mascot generic format
MHC	Major histocompatibility complex
min	Minute
MMP	Matrix metalloproteinases
MOPS	3-(*N*-Morpholino)propanesulfonic acid
MS	Mass spectrometry
MS1	Full mass spectrum
MS2	Tandem mass spectrum
MW	Molecular weight
MWCO	Molecular weight cut-off
m/z	Mass-to-charge ratio
N	Total number of replicates
NAT	N-terminal acetyltransferase
N-CLAP	N-terminalomics by chemical labeling of the α-amine of proteins

NHS	*N*-hydroxysuccinimide
NL	Normalised intensity level
NME	N-terminal methionine excision
Nrich	N-terminal peptides enrichment on the filter
N-TAILS	N-terminal amine isotopic labeling of substrates
***P*-value**	Probability value
PBS	Phosphate-buffered saline
PEG	Polyethylene glycol
pH	Potential for hydrogen
pI	Isoelectric point
PITC	Phenyl isothiocyanate
pK_a	Acid dissociation constant
PLCγ1	Phospholipase C-gamma1
PLP	Pyridoxal-5'-phosphate
PMSF	Phenylmethanesulfonyl fluoride
PPI	Protein-protein interaction
ppm	Parts per million
PQD	Pulsed Q Collision-induced Dissociation
PRINT	PRotect, INcise, Tag
PSM	Peptide-spectrum match
PTM	Post-translational modification
PTAG	Phospho-tagging
Q	Quadrupole
QqQ	Triple quadrupole
R	Resolution (mass analyser)
RP-HPLC	Reversed-phase high-performance liquid chromatography
RF	Radiofrequency
RT	Room temperature
RT	Retention time
RS	*N*-Methylpyridinium-4-carboxaldehyde
s	Second
SATA	*N*-succinimidyl *S*-acetylthioacetate
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate

SDS-PAGE	Sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SILAC	Stable isotope labeling with amino acids in cell culture
SRM	Selected reaction monitoring
STONE	Selective Transamination Of N-Ends
SUMO	Small ubiquitin-like modifier
SWATH	Sequential window acquisition of all theoretical fragment-ion spectra
TAP	Tandem affinity purification
TCA	Trichloroacetic acid
TCR	T-cell receptor
TEV	Tobacco etch virus
TFT-α	Tumor necrosis factor-α
TIC	Total ion current chromatogram
TMPP	*N*-succinimidyloxycarbonylmethyl tris(2,4,6-trimethoxyphenyl)phosphonium bromide
TMT	Tandem mass tag
TNBS	2,4,6-trinitrobenzenesulfonic acid
TOF	Time-of-flight
Tris	Tris(hydroxymethyl)aminomethane
Ub	Ubiquitin
UPS	Ubiquitin-proteasome system
UV	Ultraviolet
UVPD	Ultraviolet Photodissociation
v	Volume
w	Weight
XIC	Extracted ion chromatogram
Y2H	Yeast two-hybrid
z	Electric charge

Chapter 1. Introduction

1.1 Proteome and proteomics

After the completion of the Human Genome Project in 2003, we are now living in a post-genomic era with more than 5,000 reference genomes available to the public (source: Kyoto Encyclopedia of Genes and Genomes, 12 Jan. 2018). However, the number of protein-coding genes does not faithfully reflect the complexity of organisms. For instance, human only possesses approximately 19,600 protein-coding genes (The UniProt Consortium, 2013), which are even fewer than those in a small worm like *Caenorhabditis elegans* (20,447; Bass *et al.*, 2016). The discrepancy between gene number and species complexity drives further research on the function of the "Junk DNA", non-coding RNA, as well as proteins. As described in the central dogma of molecular biology, proteins are the receiving terminal of genetic information that flows from nucleic acids (Crick, 1970). The central dogma is often illustrated as DNA ↔ RNA → Protein. Proteins are essential biomolecules that are synthesised through the assembly of amino acids into polypeptide chains. These biomolecules participate in virtually all cellular processes and contribute in a major way to the phenotype of individual organisms.

Given the importance of protein functions and the availability of genome sequences, there have been tremendous efforts to study the entire complement of proteins expressed in an organism. The original idea of the "total protein set" is analogous to the genome in concept, and the term "proteome" was first introduced by Marc Wilkins at a meeting in Siena, 1994 to properly describe this idea (Wilkins *et al.*, 1996). Similarly, the study of the proteome is termed "proteomics" by analogy with the genomics. Even though the traditional view of the "static genome" has drastically changed in the last decade, the proteome is still considered much more dynamic than the genome, by several orders of magnitude (Brower, 2001). Within an individual organism, the proteome varies considerably between different tissue/cell types and different times. Furthermore, the proteome exhibits characteristic perturbations in response to environmental stimuli and in disease states. As a result, the proteome is currently defined as the total set of expressed proteins in a specific biological context (cell, tissue, or organism) at a particular time and under defined conditions (Chandramouli and Qian, 2009).

Proteomics is a multifaceted research field, encompassing a wide variety of technical disciplines that range from the yeast two-hybrid (Y2H) system to microarray-based

techniques (Aebersold and Mann, 2003). Such techniques can be employed to determine protein-protein interactions (PPIs), protein structures and locations, as well as to detect and measure changes in gene expression under different conditions. However, it is well acknowledged that the global analysis of proteins poses great technical challenges. The difficulty partly owes to the spatio-temporal variability and dynamic nature of the proteome (as described above). This difficulty is further compounded by alternative splicing at mRNA level, post-translational modifications (PTMs) of proteins, and PPIs. Another tier in the complexity originates from the physico-chemical diversity of proteins and the wide dynamic range of protein expression (Altelaar *et al.*, 2013).

1.1.1 Mass spectrometry-based proteomics

Over the past twenty years, there have been many technological breakthroughs in protein separation, analytical science, and bioinformatics to resolve the said complexity. In this regard, proteomics is a research field driven by technology (Wilkins and Appel, 2007). Among all the breakthroughs, the maturation of mass spectrometry (MS) has fundamentally transformed the proteomic research. MS-based proteomics provides reliable identification and quantitation of proteins in a high-throughput manner. There are a large variety of MS instruments, but each basically consists of an ion source, one or more mass analysers, and a detector (Figure 1.1). Such systems measure the mass-to-charge ratio (*m*/*z*) of charged particles from which the mass can usually be deduced.

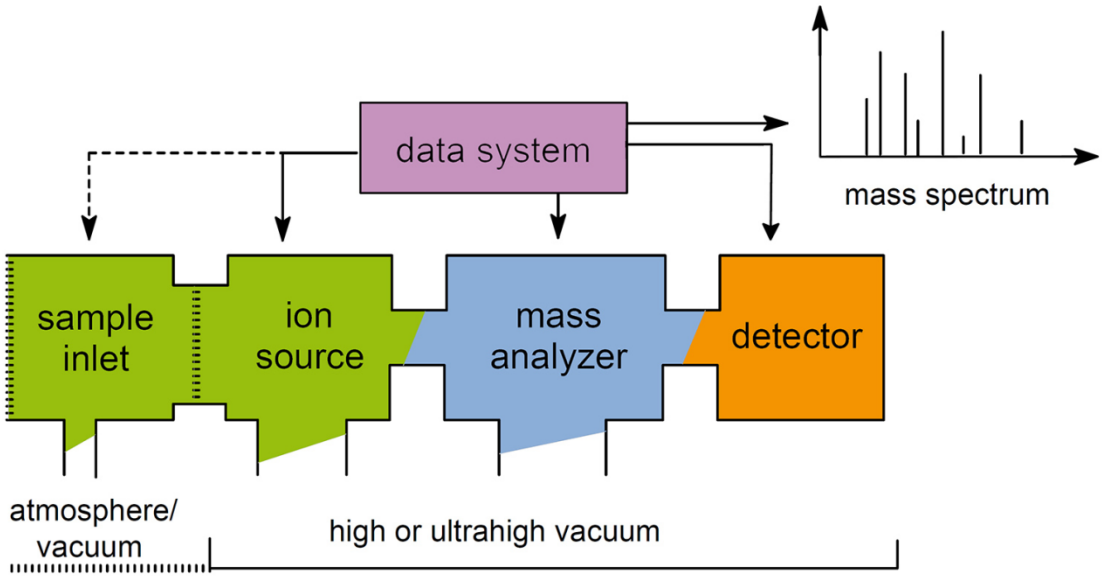


Figure 1.1 Basic setup of mass spectrometry (MS) instrumentation, which invariably consists of an ion source, one or more mass analysers, and a detector (adapted from Gross, 2017). In MS-based proteomics, protein samples are typically converted to peptides and then introduced to a liquid chromatography (LC) system. This sample inlet is coupled online to an MS instrument.

Although MS is an essential tool in proteomics, it is not a novel technology specifically developed for this purpose. The first mass spectrometer was invented by J. J. Thomson in the early 20th century to measure the *m*/*z* of atoms and indeed the electron (reviewed in Griffiths, 2008). The use of MS largely remained in the realm of physics and chemistry until the late 1980s, when the ionisation of proteins and peptides was made possible. Two "soft" ionisation techniques, matrix-assisted laser desorption/ionisation (MALDI) and electrospray ionisation (ESI), were developed to ionise such biological macromolecules for MS analysis without significant fragmentation (Karas and Hillenkamp, 1988, Fenn *et al.*, 1989). The MALDI technique involves pulsed laser irradiation of protein/peptide analytes that have been mixed with light-absorbing compounds (i.e. the "matrix"). Upon irradiation, the matrix absorbs ultraviolet (UV) light and sublimes, transferring energy to the embedded analytes for ionisation. In MALDI-MS, protein/peptide analytes are converted into gas-phase ions mostly with a single charge (Karas *et al.*, 2000). Ionised analytes are then accelerated by electric potentials into most often a time-of-flight (TOF) mass analyser.

A TOF analyser deduces the *m*/*z* values of ionised analytes by simply measuring the time taken by ions to fly through a tube of set length under vacuum to reach the mass detector. Such deduction is based on the direct relationship between the flight time and the *m*/*z* of ions, which are accelerated by application of a constant voltage. This relationship can be expressed as: $m/z = K \times t^2$, where **t** is flight time and **K** is a coefficient determined by the voltage and tube length (El-Aneed *et al.*, 2009). A reflector that turns around the ion path is employed by modern TOF analysers to correct for small differences in initial kinetic energies of the accelerated ions, as well as to increase the flight distance. As a result, TOF analysers provide outstanding resolving power (~ 25,000) and mass accuracy (5 – 10 parts per million, ppm), the fastest data acquisition, as well as a very high *m*/*z* range (Calvete, 2014).

MALDI-MS is often used to characterise protein/peptide samples of low complexity (Aebersold and Mann, 2003). Recently, there has been a tremendous advance in imaging techniques that rely on MALDI-MS (referred to as "MALDI-MS imaging"). Their applications include the resolution of protein spatial distribution in tissue specimens (Aichler and Walch, 2015).

Meanwhile, the ESI technique employs a strong electric field (≥ 2 kV) to charge liquid droplets, which emerge from the tip of an electrospray capillary and typically contain proteins or peptides previously separated by liquid chromatography (LC). The charged droplets are emitted from the capillary to the counter electrode at atmospheric pressure. Due to solvent evaporation, the charged droplets gradually shrink and then produce smaller

droplets by Coulomb explosions (Dole *et al.*, 1968). Ultimately, each charged droplet contains a single protein/peptide and yield gas-phase ions (Thomson and Iribarne, 1979). In contrast to MALDI, ESI can be conveniently coupled to LC systems as well as a large variety of mass analysers (see below). Equally importantly, it yields multiply-charged peptide and protein ions, which brings large molecules into the analysable range. ESI-MS is therefore currently preferred for analysing complex proteomes (Aebersold and Mann, 2003).

A quadrupole (Q) mass analyser is composed of four parallel magnetic rods to which a direct current (DC) and a superimposed radiofrequency (RF) potential are applied. An alternating electromagnetic field is thus created, serving as a mass filter for ions of the specified *m*/*z* to pass through and reach the mass detector. Ions other than those of the desired *m*/*z* value have unstable trajectories and will collide with the metal rods and hence are not detected. A range of *m*/*z* values can be systematically scanned by continuously varying the amplitudes of the applied DC/RF potentials at a fixed ratio (Dawson, 2013).

A Q analyser is of limited resolution (4,000) and mass accuracy (100 ppm), and lacks the ability to perform tandem mass spectrometry (MS/MS) on its own (Calvete, 2014). The MS/MS technique is pivotal to contemporary proteomics since not only does it determines the *m*/*z* value of an ionised peptide but also provides sequence information of the peptide through a fragmentation process (Steen and Mann, 2004). The lack of MS/MS capability can be compensated for, however, by coupling a Q analyser to other mass analysers. This is exemplified by triple quadrupole (QqQ) and other hybrid configurations (e.g. Q-TOF). The high dynamic range (10^7) of Q analysers confers advantages in quantitative proteomics (Hart-Smith and Blanksby, 2012).

Ion trap mass analysers, in either a linear or three-dimensional (3D) configuration, employ an RF electric field to confine ionised analytes. Trapped ions are then excited by increasing the RF amplitude and ejected to reach the mass detector outside the ion trap (Cooks *et al.*, 1991). As a result, the *m*/*z* of these ions can be scanned according to the correlation between RF amplitudes and *m*/*z* values (Wong and Cooks, 1997). The strengths of ion trap analysers primarily lie in their affordability, robustness, and simplicity. In addition, ion trap analysers possess high sensitivity and the capability to perform multistage mass spectrometry (MS^n) analysis (Hart-Smith and Blanksby, 2012). But similar to Q analysers, ion traps have shortcomings in resolving power and mass accuracy.

The Orbitrap is the most recent entry in this series of mass analysers. It is a modified ion trap consisting of a central spindle electrode and two endcap electrodes. Unlike conventional ion

traps, the Orbitrap confines ionised analytes by a ramping of electrostatic fields on the central electrode (a procedure known as "electrodynamic squeezing"; Hu *et al.*, 2005). Once trapped inside the analyser, ions of different *m*/*z* values orbit the central electrode at different frequencies and are eventually separated into discrete rings. Such oscillations induce an image current on the endcap electrodes, which is then recorded by the mass detector. The *m*/*z* of these ions can be deduced through fast Fourier transformation of their orbiting frequencies (Lange *et al.*, 2014a). Compared with other mass analysers described above, the Orbitrap exhibits the highest resolution (1,000,000) and accuracy (2 – 5 ppm), which underlie its widespread use in proteomics (Denisov *et al.*, 2012, Calvete, 2014).

As stated before, once a peptide is ionised and its *m*/*z* value (and thus the mass) determined, sequence information can be further provided by MS/MS, although this technique has not yet achieved the same extent of success when analysing intact proteins, i.e. “top-down proteomics” (Catherman *et al.*, 2014). In MS/MS, an ionised peptide is selectively fragmented into its basic constituents and the resulting mass spectrum is interpreted to determine the peptide sequence. There are two operation modes for MS/MS: tandem-in-space and tandem-in-time. The former refers to the physical coupling of two mass analysers for determination of mass and sequence information, respectively (de Hoffmann and Stroobant, 2007). The two analysers are separated by a device for peptide fragmentation, known as a collision cell, as exemplified by QqQ and Q-TOF configurations. On the other hand, both mass and sequence information of a peptide can be determined using a single mass analyser in the tandem-in-time mode. This is only possible with analysers that possess trapping capabilities (Johnson *et al.*, 1990).

The fragmentation step is critical to the success of MS/MS experiments, and there are a plethora of fragmentation techniques currently available to proteomics researchers. These include Collision-induced Dissociation (CID), Higher-energy Collisional Dissociation (HCD), Electron Transfer Dissociation (ETD), Pulsed Q Collision-induced Dissociation (PQD), and UV Photodissociation (UVPD). These fragmentation techniques ultimately define the resulting ion species, and each technique possesses unique characteristics that merit its specific application in MS/MS (Sleno and Volmer, 2004).

As the most universal technique, CID has been employed for peptide fragmentation since the 1980s (Biemann, 1986, Hunt *et al.*, 1986). During this process, ions that correspond to a peptide of interest (i.e. “precursor ions”) are first transmitted to a collision cell in the case of a QqQ configuration, or confined in an ion trap analyser whilst other ions are resonantly ejected from the trap. In either case, the selected precursor ions collide with inert gases (He,

Ar, or N_2) and kinetic energies are imparted onto the precursor ions. Such events increase the internal energy of precursor ions, inducing fragmentation mainly through the lowest energy pathways (i.e. cleavage of peptide bonds; Roepstorff and Fohlman, 1984). The wide acceptance of CID in MS-based proteomics is primarily attributed to the ease of implementation and high fragmentation speed/efficiency (Wells and McLuckey, 2005).

In peptide fragmentation along its backbone, breakage of C_α–C, C–N (the peptide bond), or N-C_α bond is possible between any two adjacent amino acid residues. Each bond breakage gives rise to a neutral fragment and a charged species (i.e. a "fragment ion"). Depending on which bond is cleaved, a charged species is designated as *a*-, *b*-, or *c*-ion if the charge is retained on the amino (N)-terminal fragment. On the other hand, an *x*-, *y*-, or *z*-ion is produced if the charge is retained on the carboxyl (C)-terminal fragment (Figure 1.2; Roepstorff and Fohlman, 1984, Johnson and Biemann, 1989). With respect to tryptic peptide fragmentation by CID, the charged fragments are predominantly *b*- and *y*-ion species (Wysocki *et al.*, 2000). This knowledge of fragmentation patterns greatly facilitates interpretation of the resulting mass spectra. However, CID has intrinsic limitations in detecting fragment ions in the low *m*/*z* range as well as in preserving labile chemical groups on peptides subjected to post-translational modifications (PTMs; Sleno and Volmer, 2004, Creese and Cooper, 2007). These limitations largely hinder the use of CID for isobaric tag-based quantitation or PTM identification (see sections 1.1.2 and 1.2).

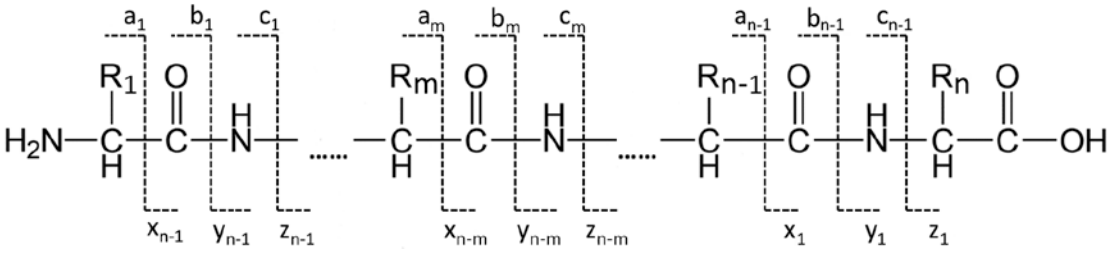


Figure 1.2 Roepstorff's nomenclature of peptide fragmentation in tandem mass spectrometry (MS/MS; modified from Qi and Volmer, 2016). Cleavage of C_α–C bonds along the peptide backbone gives rise to *a*- and *x*-ions corresponding to N- and C-terminal fragments, respectively. Similarly, fragmentation of C–N bonds (i.e. the peptide bonds) produces *b*- and *y*-series fragment ions, whereas *c*- and *z*-ions result from fragmentation of N-C_α bonds along the peptide backbone. For a peptide of length **n**, *a*-, *b*-, and *c*-ions are labelled consecutively from the N-terminus as a_m, b_m, and c_m, where the subscript **m** denotes the number of amino acid residues in each fragment ion. The complementary *x*-, *y*-, and *z*-ions are designated as $x_{(n-m)}$, $y_{(n-m)}$, and $z_{(n-m)}$, where **n-m** equals the number of residues each fragment ions contains.

Alternative techniques have therefore been developed to complement CID for peptide fragmentation. Two such techniques of particular relevance, HCD and ETD, are introduced here. HCD is a variant of CID only available to mass spectrometers equipped with the Orbitrap analyser. In such instruments, external to the Orbitrap are a multipole collision chamber (i.e. the HCD cell) and an ion storage device (i.e. the C-trap) where beam-type CID fragmentation and trapping of fragment ions by a high RF voltage (2.5 kV) take place, respectively (Olsen *et al.*, 2007). The trapped fragment ions are then injected into the Orbitrap analyser for spectral acquisition at high resolution. Compared with CID, HCD also produces *b*- and *y*-series fragment ions but the fragmentation pattern is slightly different (de Graaf *et al.*, 2011). More importantly, HCD overcomes the intrinsic limitation of CID in low *m*/*z* regions and is thus compatible with isobaric tagging (Bantscheff *et al.*, 2008).

Developed by Syka *et al.* (2004), ETD is a fragmentation technique based on the chemistry of ion/ion reactions (McLuckey and Huang, 2009). In this process, electrons are transferred from reagent anions (e.g. fluoranthene) to precursor ions at higher charge states (e.g. +3). The electron transfer converts precursor ions into radical species, and a rearrangement of the radical species ultimately induces breakage of backbone N-C_α bonds in the case of peptide fragmentation. Consequently, ETD typically produces *c*- and *z*-series fragment ions from tryptic peptides, in contrast to *b*- and *y*-ions present in CID or HCD spectra (reviewed in Zhurov *et al.*, 2013).

Another distinction of ETD is the preservation of labile PTMs (e.g. phosphorylation and glycosylation) during peptide fragmentation (Mikesh *et al.*, 2006). ETD is therefore well suited to the identification of exact PTM sites within a modified peptide. However, the fragmentation efficiency of ETD is lower than that of CID, leading to fewer total peptide identifications (Zubarev *et al.*, 2008, Guthals and Bandeira, 2012). Consequently, ETD often serves as a complementary technique to improve peptide identification by CID or HCD.

With the advent of all the techniques mentioned above, contemporary MS-based proteomics typically uses a “bottom-up” strategy and consists of three stages (Figure 1.3). The three stages are: I. sample preparation: proteins are extracted from cell/tissue samples and then digested into peptides with a site-specific protease (e.g. trypsin). II. LC-MS/MS analysis: peptides are separated by LC and then ionised through ESI; a mass spectrometer records the *m*/*z* of peptide ions (full mass spectra, MS1), which are then selected for fragmentation and acquisition of the fragment-ion spectra (tandem mass spectra, MS2). III. data analysis: the combination of MS1 and MS2 data facilitates protein identification and quantitation using specialised software (e.g. Mascot, MaxQuant, and Skyline).

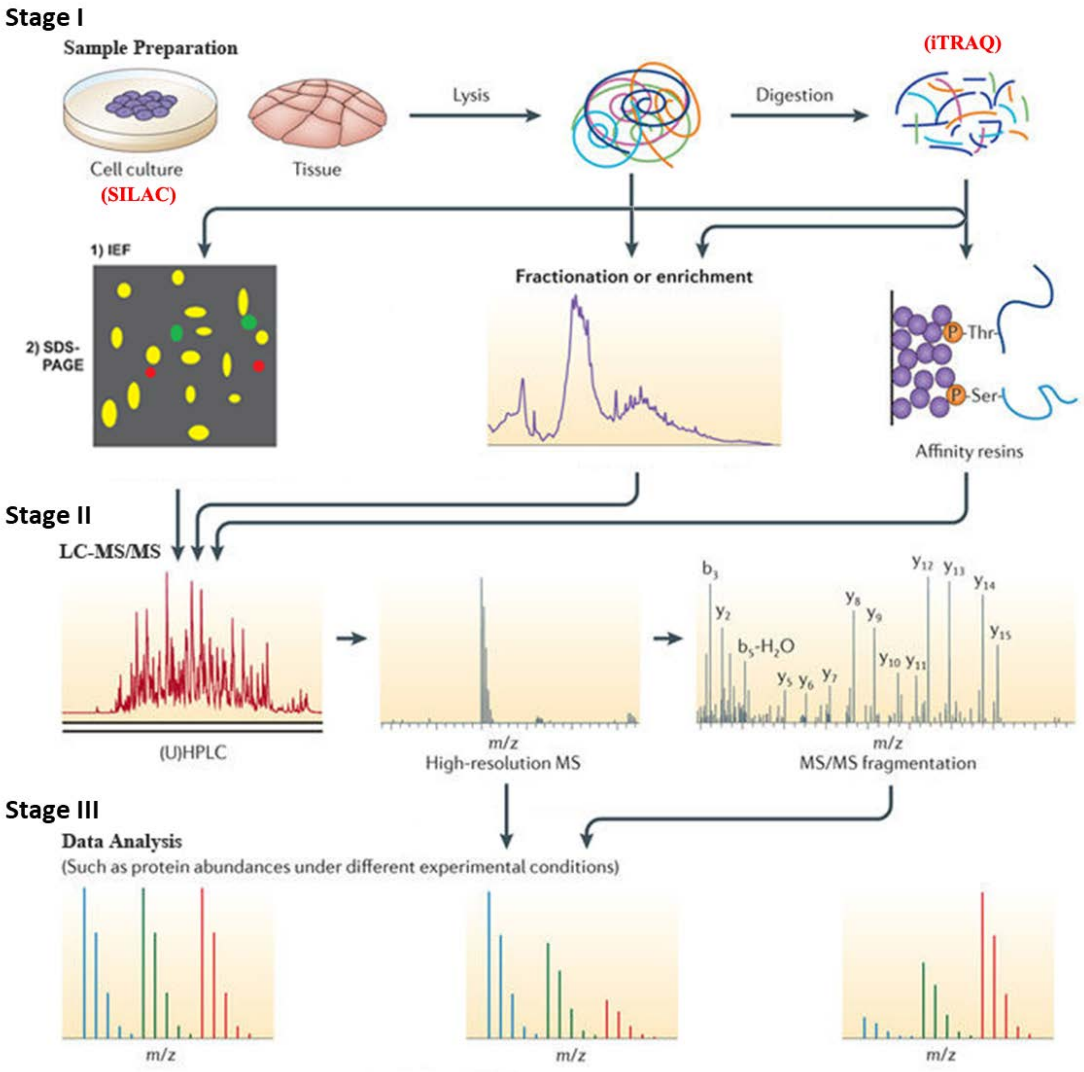


Figure 1.3 Overview of MS-based, bottom-up proteomics. Typical workflow consists of three stages: I. sample preparation; II. liquid chromatography–tandem mass spectrometry (LC-MS/MS); III. data analysis. Methods of label-based quantitation are denoted in red (adapted from Altelaar *et al.*, 2013).

In Stage I, protein extraction and digestion have often been intertwined with protein separation by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) or two-dimensional (2D) gel electrophoresis, i.e. isoelectric focusing (IEF) + SDS-PAGE. Alternatively, sample fractionation (e.g. reversed-phase or ion-exchange chromatography) or affinity purification (AP) is carried out at protein or peptide level to reduce proteome complexity. In addition, protein or peptide samples are often labelled with stable isotopes via SILAC (stable isotope labeling with amino acids in cell culture) for protein quantitation (Ong *et al.*, 2002). Alternatively, this can be achieved using isobaric tags, e.g. iTRAQ (isobaric peptide tags for relative and absolute quantification; Ross *et al.*, 2004) or TMT (tandem mass tags; Thompson *et al.*, 2003).

As reviewed by Aebersold and Mann (2016), three types of bottom-up methods can be adopted in Stage II: data-dependent acquisition (DDA), targeted proteomics (e.g. selected reaction monitoring, SRM), and data-independent acquisition (DIA) that comprises several cutting-edge techniques. The classical DDA method allows high-throughput protein identification (referred to as "shotgun proteomics"), hence it is primarily used for discovery purposes. In DDA, the most abundant few peptide precursor ions surveyed in MS1 scans are isolated for fragmentation according to their signal intensities (Domon and Aebersold, 2006). The *m*/*z* of the resulting fragment ions are then recorded in MS2 scans. The aforesaid software tools can perform a match between experimental MS2 spectra and the theoretical ones generated from sequence databases (e.g. Swiss-Prot; Bairoch and Apweiler, 2000). This process allows peptides to be identified and the corresponding proteins inferred.

Compared with DDA, targeted proteomics requires *a priori* knowledge of the proteins of interest: specifically, one must know the *m*/*z* values of peptide precursors that are representative of the target proteins, and the *m*/*z* values of corresponding fragment ions that have high signal intensity (Lange *et al.*, 2008). Several precursor/fragment ion pairs, or "transitions", are selectively monitored over LC retention time by SRM-type techniques. This ensures the specificity, sensitivity, and high dynamic range required for accurate and reproducible quantitation of target proteins (Picotti and Aebersold, 2012). In targeted proteomics, data analysis can be performed with the Skyline software (MacLean *et al.*, 2010).

There is also a clear difference in the way that mass spectrometers are operated between DDA and the emerging DIA method. The history of DIA can be traced back to the development of MS^E, a technique that utilises alternating low and high collision energy for simultaneous acquisition of MS spectra (Purvine *et al.*, 2003, Plumb *et al.*, 2006). MS^E is similar to DDA in providing mass information on both precursor and fragment ions, but the acquisition of fragment-ion spectra is independent of the signal intensity of precursor ions.

The concept of DIA is, however, best exemplified by the SWATH-MS (sequential acquisition of all theoretical mass spectra) technique. It cycles through a defined *m*/*z* range, which is divided into several precursor isolation windows (e.g. 25 *m*/*z* units each), over LC retention time (Gillet *et al.*, 2012). The SWATH-MS technique simultaneously fragments all peptide precursor ions in each *m*/*z* window and records the fragment-ion spectra in a stepwise and iterative fashion. SWATH-MS (and similar DIA techniques) can generate time-resolved data of fragment ions for all detectable peptides in a sample, thus overcoming the stochasticity and sampling bias typically associated with DDA (Michalski *et al.*, 2011).

In theory, DIA combines the advantages of both DDA and targeted proteomics for high-throughput protein identification and reproducible quantitation. However, challenges in deciphering DIA data have been well established: in DIA data, each mass spectrum is highly convoluted due to the presence of numerous fragment ions that are derived from different peptide precursors within the same *m*/*z* window; the correspondence between precursor ions and their fragments is thus much more difficult to ascertain than in DDA data analysis (Doerr, 2014). At present, DIA data are analysed primarily through targeted extraction approaches, which rely on spectral libraries built from previous DDA/SRM experiments (Gillet *et al.*, 2012). Similar to the analysis of SRM data, targeted extraction approaches involve construction of a peptide transition list from spectral libraries, generation of extracted ion chromatograms (XICs) for the selected fragment ions, scoring peptide groups by their chromatographic and signal-intensity profiles, as well as statistical validation based on decoy transitions (reviewed in Bilbao *et al.*, 2015). OpenSWATH and Spectronaut represent software tools that can automate this process for high-throughput analysis (Rost *et al.*, 2014, Bruderer *et al.*, 2015).

The dependence of targeted extraction approaches on spectral libraries is a limiting factor in DIA data analysis in spite of their popularity, and library-free approaches have thus been developed to alleviate this limitation (Tsou *et al.*, 2015). Such approaches directly generate pseudo-MS2 spectra from DIA data without the need for prior knowledge. The pseudo-MS2 spectra are then utilised for the sequence database search similar to that in the analysis of DDA data. Recently, DIA-Umpire has come into prominence in this category owing to its versatility and cross-platform applications (Tsou *et al.*, 2016). However, questions are often raised regarding the reliability of library-free approaches since DIA data are highly complex (Rost *et al.*, 2014). In conclusion, both targeted extraction and library-free approaches require further developments to realise the full potential of DIA.

Typically, targeted proteomics and DIA require specific MS instruments. For instance, SRM-type techniques are mainly compatible with QqQ instruments, whereas MS^E and SWATH-MS are limited to Waters and AB SCIEX Q-TOF platforms, respectively. In contrast, most of the aforementioned mass analysers and their hybrids can be used for DDA. As the dominant method in shotgun proteomics, DDA is constantly improving and enabling scientific breakthroughs in many aspects (see section 1.1.2). At Xi'an Jiaotong-Liverpool University (XJTLU), an Orbitrap Elite™ hybrid ion trap-Orbitrap mass spectrometer (Thermo Fisher Scientific) is typically operated in DDA mode for proteome discovery.

1.1.2 Major focus areas of MS-based proteomics

With the numerous technological breakthroughs, MS-based proteomics has become an indispensable approach to study the correlation between the genotype and the phenotype (and its variations). Currently, the discovery of protein biomarkers is one of the most vibrant fields in proteomics. Proteins of which the abundance consistently correlates with a disease phenotype are considered potential biomarkers with paramount diagnostic or prognostic values (Altelaar *et al.*, 2013). Searching with "protein biomarker" as key words now yields more than 600,000 publications in the PubMed database (Figure 1.4).

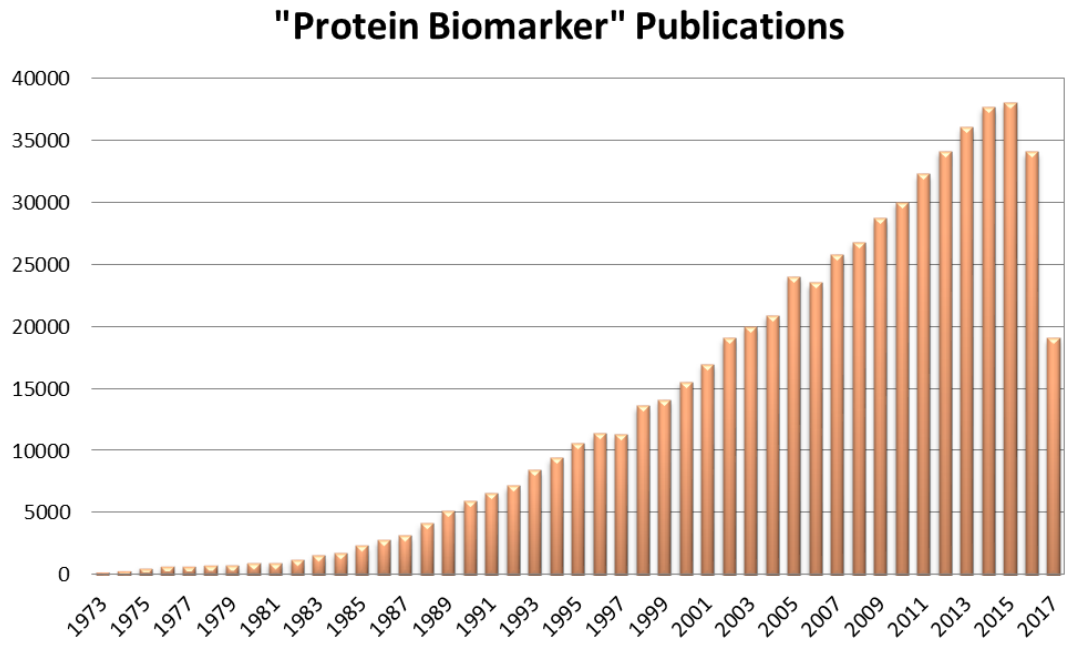


Figure 1.4 Protein biomarker discovery is a dramatically expanding field in biological research, reflected by the more than 600,000 publications (source: PubMed, 12 Jan. 2018). PubMed was searched using the term: "protein biomarker".

Apart from such clinical applications, MS-based proteomics is heavily applied to study the following four aspects of the proteome: protein identification, protein quantitation, PPIs, and protein PTMs. Protein identification is mainly performed using the DDA approach, which allows for hypothesis-free profiling of the proteome of interest. With improvements in sample preparation and LC-MS/MS instrumentation, the yeast proteome can now be analysed in one hour (h) with > 90 % coverage (Hebert *et al.*, 2014). Similarly, collective efforts to profile the entire human proteome (i.e. the Human Proteome Project) have led to the identification of more than 17,000 proteins, accounting for 87 % of the predicted human proteome (Wilhelm *et al.*, 2014, Kim *et al.*, 2014).

The last two decades have witnessed MS-based proteomics gradually turning quantitative. Many protein quantitation methods have been devised so far, allowing MS-based proteomics to measure the differential expression of proteins (e.g. as a function of cellular or disease states), or to determine the absolute quantity of distinct proteins within a sample (Ong and Mann, 2005). Quantitative proteomics is often divided into label-based and label-free strategies. As mentioned above, SILAC and iTRAQ/TMT belong to the label-based strategies. They are widely employed for relative quantitation on the basis that native peptides and their labelled counterparts behave similarly in LC-MS/MS (Bantscheff *et al.*, 2007).

Briefly, SILAC entails the *in vivo* metabolic labelling of proteins via cell culture in the presence of specific "heavy" (i.e. ^{13}C and/or ^{15}N-labelled) amino acids, e.g. lysine (K) and arginine (R). In such experiments, the control sample is typically heavy-labelled and then 1:1 mixed with the test sample where cells have been grown with "light" (i.e. unlabelled) amino acids. Relative quantitation is achieved by determining the ratio of light/heavy peptide pairs, which exhibit a defined mass shift in MS1 spectra (Ong *et al.*, 2002).

On the other hand, iTRAQ/TMT involves the *in vitro* labelling of proteolytic peptides with isobaric mass tags. In other words, a chemical reaction is carried out after protein extraction and protease digestion, which results in the attachment of a mass tag at the N-terminus and K residues of each peptide. In an iTRAQ/TMT experiment, each sample is labelled with a variant of the mass tag; the differentially labelled samples are then pooled and analysed simultaneously by LC-MS/MS (Thompson *et al.*, 2003, Ross *et al.*, 2004). Due to the isobaric nature, the same peptides from differentially labelled samples will have identical *m*/*z* values and thus appear as a single composite peak in MS1 spectra. Fragmentation of such precursor ions then releases a reporter ion, with a characteristic *m*/*z* value, from each isobaric mass tag. The number of reporter ions released and detected is proportional to the relative abundance of the peptide in question, which in turn reflects the relative abundance of the protein from which it is derived. Therefore, the same peptides from different samples can be distinguished and relatively quantified according to the *m*/*z* and intensities of the reporter ions in MS2 spectra, respectively.

As described above, SILAC and iTRAQ/TMT adopt different strategies for relative quantitation: the former relies on MS1 spectra whereas the latter is MS2-based. Compared with iTRAQ/TMT, the incorporation of stable isotopes and sample pooling are both performed at earlier stages in an SILAC experiment. This helps to diminish technical variations and thus ensures higher quantitation accuracy. Conversely, SILAC is more

expensive and less well-suited to multiplexing, which is the primary strength of the isobaric tagging methods (reviewed in Bantscheff *et al.*, 2012). In addition to relative quantitation, label-based strategies have also been employed for absolute quantitation. For instance, isotope-labelled peptides that correspond to tryptic peptides derived from target proteins, known as AQUA peptides (AQUA stands for "absolute quantification"), are spiked into a sample in a known concentration (Gerber *et al.*, 2003). These AQUA peptides serve as internal standards to determine the exact concentration of their unlabelled counterparts, which are derived from the target proteins in the sample. At present, AQUA peptides can be obtained commercially albeit at a relatively high cost. This method is thus better suited to targeted analysis in lieu of shotgun proteomics.

The throughput of quantitative proteomics has increased dramatically since the advent of label-free strategies, which historically focused on relative quantitation of shotgun proteomic data (Old *et al.*, 2005). In general, label-free strategies can be subdivided into two categories: the first counts the number of peptide-spectrum matches (PSM) or the frequency of the observed peptides for each identified protein (i.e. "spectral counting"), whereas the second extracts peptide signal intensities from LC-MS/MS data in order to infer protein abundance (i.e. intensity-based methods). The spectral counting methods are based on the assumption that the abundance of a protein correlates with the number of PSMs detected per protein (Washburn *et al.*, 2001). On the other hand, the linear correlation between the signal response and the abundance of a peptide over four orders of magnitude in LC-MS/MS lays the foundation for the intensity-based methods (Chelius and Bondarenko, 2002). It should be stressed though that rigorous normalisation is essential to the accurate quantitation by label-free strategies, which are adversely affected by both biological and technical variations (Bantscheff *et al.*, 2012). Additionally, both types of label-free quantitation strategy generally only work well with abundant proteins.

Although label-based strategies (e.g. AQUA) are still regarded as superior for absolute quantitation, label-free strategies have gained popularity in recent years. The primary strength of such methods lies in the fact that they can be applied, even retrospectively, to large cohorts of shotgun proteomic data (Old *et al.*, 2005). Two spectral counting methods, APEX (absolute protein expression; Lu *et al.*, 2007) and emPAI (exponentially modified protein abundance index; Ishihama *et al.*, 2005), are described here. APEX employs a sophisticated classification algorithm to estimate the detection probability of unique peptides for a given protein. For each protein, a correction factor $\mathbf{O}_i$ is then derived from the estimated detection probability and in turn facilitates absolute quantitation by normalising

the number of observed PSMs per protein. Meanwhile, the emPAI method assigns each protein with an abundance estimate (calculated as $10^{PAI} - 1$, where **PAI** is defined as the number of observed peptides per protein divided by the corresponding number of theoretically observable peptides). The emPAI value for each identified protein can be directly retrieved from the results of database searches using Mascot (Perkins *et al.*, 1999).

Similar to spectral counting, the intensity-based methods have also been adopted in efforts to achieve absolute quantitation. For instance, an iBAQ (intensity-based absolute quantification) value is calculated by dividing the sum of all observed peptide intensities for a protein by the corresponding number of theoretically observable peptides (Schwanhausser *et al.*, 2011). Unlabelled spike-in standards often serve the purpose of sample-specific calibration, which facilitates accurate absolute quantitation. The iBAQ method has been integrated into the MaxQuant software (Cox and Mann, 2008).

MaxQuant also employs the "proteomic ruler" method as an alternative for absolute quantitation (Wisniewski *et al.*, 2014). In this method, histone proteins are set as a standard to estimate the copy number of individual proteins per eukaryotic cell without the need for spike-in standards or cell counting. The use of histone proteins is based on two assumptions: I. the abundance of total histones in an eukaryotic cell is approximately equal to the cell's DNA content, which is a constant determined primarily by the species (van Holde, 1989); and II. the ratio of protein abundances between a specific protein and total histones is approximately equal to the ratio of the summed peptide intensities between the two (Wisniewski *et al.*, 2012). The validity of proteomic ruler has been thoroughly evaluated on different species and cell types (reviewed in Wisniewski, 2017). Clearly, such an approach is limited to eukaryotic organisms as prokaryotes generally lack histones.

APEX, emPAI, iBAQ, and the proteomic ruler approaches are only a small fraction of label-free strategies developed thus far. Apart from these DDA-centric methods, targeted proteomics (e.g. multiplexed SRM) and DIA (e.g. SWATH-MS) are increasingly useful for label-free quantitation across multiple samples. This is exemplified by the quantitation of 192 proteins from mouse liver samples in two metabolic states using SRM (Wu *et al.*, 2014). In comparison, > 2,600 proteins were quantified by SWATH-MS across 386 mouse liver samples with the same experimental design (Williams *et al.*, 2016).

Over the years, proteomics has extended beyond the identification and quantitation of individual proteins, to determining protein interaction networks and the structures of macromolecular assemblies. The interconnectivity of proteins into complexes forms the

basis of many cellular functions and hence largely determines the phenotype of a cell (Altelaar *et al.*, 2013). Affinity purification–mass spectrometry (AP-MS) has been extensively applied to identify PPIs, complementing the conventional Y2H assay. AP-MS employs antibodies to purify the endogenous "bait" protein and its binding partner ("prey") for high-throughput identification (Dunham *et al.*, 2012). Alternatively, bait proteins are affinity tagged in a carefully controlled manner so that the tagged proteins are expressed at a physiological level. Tandem affinity purification (TAP) is an improvement in AP-MS that employs two sequential AP steps to significantly reduce background binding (Rigaut *et al.*, 1999). As TAP is less efficient in multicellular organisms, Rees *et al.* (2011) took advantage of transposable elements to achieve parallel tagging and AP-MS in fruit flies and other complex organisms. Quantitative methods (e.g. SILAC) allow AP-MS to further determine the stoichiometry of PPIs, distinguishing stable interactions from the weaker, transient ones (Rees *et al.*, 2015).

Multiple MS-based techniques have been applied to complement X-ray crystallography and nuclear magnetic resonance (NMR) for analysing the structure of large protein complexes. These include native MS, hydrogen-deuterium exchange (HDX) MS, and protein cross-linking (XL) coupled with MS. These methods are often combined with cryo-electron microscopy (cryo-EM) in an integrative approach to shed light on the topology, conformational changes, and stoichiometry of large protein assemblies (Holding, 2015). Finally, MS-based proteomics is widely used to study protein PTMs, as they result in defined shifts in protein or peptide mass. The following section will describe different types of protein PTMs and the proteomic strategies to study them.

1.2 Protein PTMs and N-terminal PTMs

1.2.1 Protein PTMs

Protein PTMs involve either the covalent attachment of chemical groups to specific amino acid residues of a protein, or the proteolysis of the protein (Walsh *et al.*, 2005). With regard to the former PTM, the chemical groups that are attached may be quite large and include oligosaccharides, lipids, and small proteins. Such attachments may be either reversible or irreversible. The latter PTM is tightly controlled and often referred to as proteolytic processing (Rogers and Overall, 2013). It is well acknowledged that cells utilise PTMs to regulate protein stability, subcellular localisation, functional state, and interaction with other molecules. As a result, cells can rapidly adapt to developmental and environmental changes. Currently, more than 400 types of PTMs have been identified, including

phosphorylation, methylation, acetylation, ubiquitination, glycosylation, as well as proteolytic processing (Giglione *et al.*, 2015).

The most notable example of PTMs is phosphorylation, which is crucial in cell signalling pathways (reviewed in Pawson and Scott, 2005). In such processes, a cascade of protein receptors and enzymes are activated or deactivated via the reversible addition or removal of phosphate groups, catalysed by opposing enzymes (protein kinases or phosphatases, respectively). These phosphorylation events mediate signal transduction within the cell, often leading to changes in gene expression that direct cellular responses to external stimuli. Conversely, dysregulated protein phosphorylation is frequently associated with different diseases including cancer (Olsen and Mann, 2013).

Western blotting with specific antibodies is by convention the "gold standard" method to detect the presence of target proteins and protein PTMs (Aebersold *et al.*, 2013). However, issues are often raised regarding the specificity of the antibodies targeting PTMs and the variability in the results (Campa and Kypta, 2011). In this regard, MS-based proteomics is ideal for studying PTMs. It allows the determination of such modifications and their exact locations on a system-wide scale. Tremendous challenges are associated with PTM analysis, owing in large part to the complexity, dynamic regulation, and varied stoichiometry of these modifications (Altelaar *et al.*, 2013). But these challenges have been partially counteracted by the technological advances in sample preparation, MS instrumentation and quantitative analysis. Such improvements include the use of AP for a variety of PTMs (e.g. phosphorylation and ubiquitination), ETD fragmentation that facilitates identification of labile PTMs, and the increasing use of SRM and SWATH-MS. The following section is devoted to describe several representative PTMs.

1.2.2 Examples of covalent PTMs

As stated above, protein phosphorylation is by far the most extensively analysed PTM due to its functional importance in regulating not only cell signalling, but also cell cycle progression, cell growth, and apoptosis (Ardito *et al.*, 2017). Protein phosphorylation involves specific addition of a phosphate group primarily to the R group of serine (S), threonine (T), or tyrosine (Y) residues of a protein. In mammalian cells, the abundance distribution of pS/T/Y is 86.4 %, 11.8 %, and 1.8 %, respectively (Olsen *et al.*, 2006). The dramatic difference in the abundance necessitates different strategies to study them at the proteome level.

Phosphoproteomics typically involves AP and fractionation of phosphopeptides that are produced by protease digestion of phosphoproteins, followed by MS-based identification and quantitation (Figure 1.5). The first major breakthrough in this field was reported by Olsen *et al.* (2006). It employed titanium dioxide (TiO_2) and strong cation exchange (SCX) chromatography to enrich and fractionate peptides that bear pS/T, leading to the detection of 6,600 phosphorylation sites. This study also used SILAC to measure the dynamic change of phosphorylation upon growth factor stimulation. The enrichment of peptides bearing pS/T can also be achieved using immobilised metal-ion affinity chromatography (IMAC; Andersson and Porath, 1986). However, peptides that bear pY require an antibody-based approach for AP (Rush *et al.*, 2005). Such techniques have been perfected over the years, and phosphoproteomics can now detect an astonishing number of phosphorylation sites (50,000) in just a single cell line (Sharma *et al.*, 2014). Since then, the focus of phosphoproteomics has shifted towards understanding the functional aspect of this PTM, and clinical analyses of tissue samples (von Stechow *et al.*, 2015).

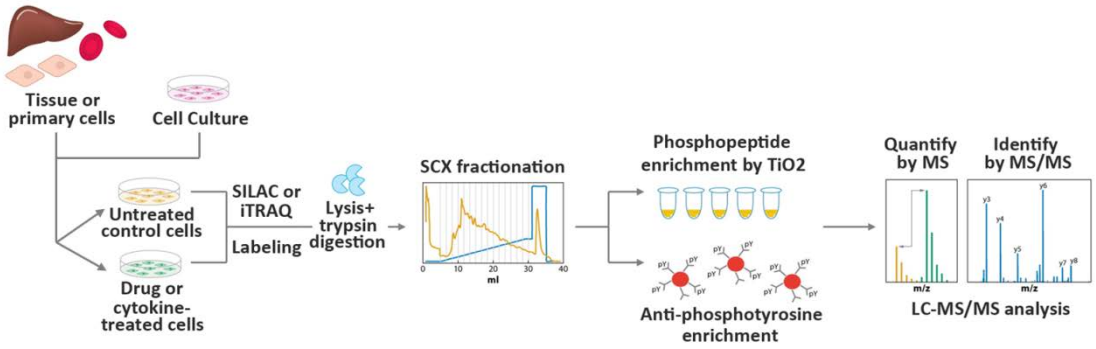


Figure 1.5 Overview of phosphoproteomics workflow (adapted from Macek *et al.*, 2009). SILAC: stable isotope labeling by amino acids in cell culture; iTRAQ: isobaric tags for relative and absolute quantitation; SCX: strong cation exchange; TiO_2: titanium dioxide.

The importance of AP is also highlighted by proteomic studies on protein ubiquitination and another PTM involving small ubiquitin-like modifier (SUMO) proteins, i.e. SUMOylation. The former PTM starts with the formation of an isopeptide bond between the C-terminal glycine (G) of ubiquitin (Ub) and the ε-amino group of a K residue in the substrates, including other Ub proteins (Walsh *et al.*, 2005). SUMO proteins resemble Ub in protein structure and the way they are attached to target proteins as well as themselves (Gill, 2004). Protein ubiquitination elicits distinct cellular outcomes through different signalling pathways, most notably the proteasomal degradation (i.e. ubiquitin-proteasome system, UPS). This system is responsible for 80 – 90 % of cellular proteolysis in eukaryotes (Rock *et al.*, 1994). In contrast,

SUMOylation is largely confined to the nucleus, and one of its functions is to regulate DNA damage response (Hendriks and Vertegaal, 2016).

Due to the low stoichiometry, rapid degradation, and the action by opposing enzymes (i.e. deubiquitinases or SUMO proteases), protein ubiquitination and SUMOylation present enormous challenges to proteomic studies (Peng and Gygi, 2001, Ordureau *et al.*, 2015). Nonetheless, the development of K-ε-GG antibody has significantly transformed these research fields, since this antibody can recognise a GG signature in Ub/SUMO-modified proteins after protease digestion (Xu *et al.*, 2010). In the latter case, SUMO mutants are expressed in transgenic cells to bear a GG motif (Impens *et al.*, 2014). By enriching Ub/SUMO-modified proteins, K-ε-GG antibody has allowed AP-MS studies to identify 20,000 ubiquitination sites and 5,000 SUMOylation sites, respectively (Udeshi *et al.*, 2013, Hendriks and Vertegaal, 2016).

1.2.3 Proteolysis as a PTM

Proteolysis (or proteolytic processing) is not only a cellular process to degrade proteins that are marked by ubiquitination, but is also an important PTM in its own right. It involves the irreversible hydrolysis of peptide bonds, thereby dissociating proteins into smaller peptides or amino acids. This process is catalysed by a family of enzymes called proteases. With > 550 members, proteases represent one of the largest enzyme families in humans (Vidmar *et al.*, 2017). These enzymes function to regulate a large variety of physiological processes, including the immune response, cell proliferation, and apoptosis (Sanman and Bogyo, 2014). For instance, a cascade of proteolytic processing events are involved in protein maturation. These include N-terminal methionine excision (NME), removal of signal or transit peptide, cleavage of polyproteins, and the removal of precursor domains. Proteolytic processing is a tightly regulated PTM due to its irreversible nature. Failure to do so can result in many pathological conditions such as inflammation, cancer and arthritis (Turk *et al.*, 2012). To fully comprehend the cellular outcome of proteolytic processing, it is crucial to identify protease substrates, to determine cleavage specificities, and to measure protease activity and its dynamic changes under different conditions.

Contemporary protease studies employ a wide range of techniques, including substrate phage display, synthetic peptide library, substrate- or activity-based probing, and gel-based methods. In substrate phage display, fusion proteins expressed by bacteriophages contain an affinity tag and a randomised peptide, which serves as a candidate of protease substrates. This method allows rapid survey of protease cleavage specificity (Matthews and Wells, 1993).

Similarly, libraries of synthetic peptides and peptide microarrays can be treated with a specific protease in order to determine cleavage sequence specificity (reviewed in Rogers and Overall, 2013). Meanwhile, substrate- and activity-based probes serve to measure protease activity *in vitro*. These probes replace the natural substrate for a target protease, and generate fluorescent signals upon cleavage or binding by the protease (reviewed in Sanman and Bogyo, 2014).

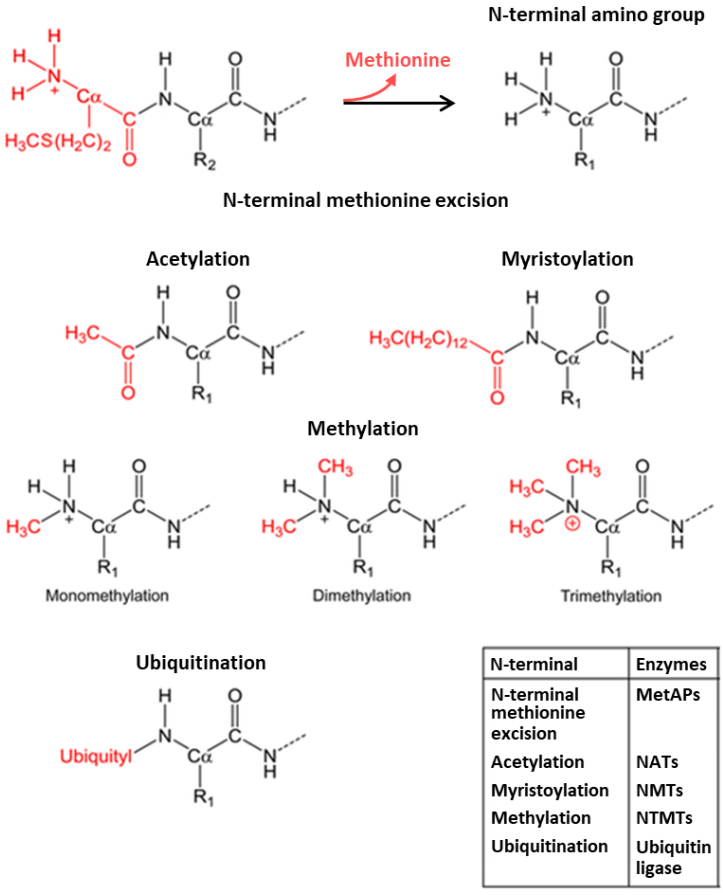
2D gel electrophoresis represents the gel-based methods to identify protease substrates on a large scale. As reviewed by Agard and Wells (2009), proteins that are digested due to *in vivo* or *in vitro* protease activity can be resolved by either IEF + SDS-PAGE or two consecutive runs of SDS-PAGE to detect protease substrates. One example is the identification of proteolytic processing events, protease substrates, and cleavage specificity in *Salmonella*, due to the activation of PhoP-PhoQ virulence regulatory system (Adams *et al.*, 1999). These early studies have spawned another gel-based method, namely PROTOMAP (protein topography and migration analysis platform). This method employs a single run of SDS-PAGE to compare proteins before and after protease digestion, and changes in protein band intensity or migration pattern are indicative of protease substrates (Dix *et al.*, 2008). All these techniques require in-gel digestion and MS analysis to determine the identity of protease substrates and protease cleavage specificity.

Similar to other PTMs, there has been a surge to study proteolytic processing (i.e. degradomics) by gel-free proteomics. However, enrichment of proteolytic products is a challenging task. The reason is as follows: proteolytic processing gives rise to newly formed N- and C-termini (i.e. *neo*-termini); these *neo*-termini on their own are chemically indistinguishable from the original N- and C-termini, thus preventing their AP (Agard and Wells, 2009). Positional proteomics (or more specifically, terminus proteomics), which labels the N- or C-terminus of a protein through chemical or enzymatic modifications, is critical to the success of studies on proteolytic processing. A subfield of positional proteomics that targets the N-terminus is named N-terminalomics, which will be described in depth in section 1.3.

1.2.4 N-terminal PTMs

As the name implies, N-terminalomics combines chemical or enzymatic labelling with MS analysis to enrich and globally analyse protein N-termini. It involves the identification of protein N-termini as well as their co- or post-translational modifications. N-terminal residue of a protein is linked to its *in vivo* stability, referred to as the "N-end rule" (Varshavsky, 2011).

In eukaryotes, the N-end rule pathway is part of UPS for protein degradation. Importantly, N-terminal PTMs are now recognised as major constituents of the N-end rule pathway, regulating the half-lives of N-terminally modified proteins. In fact, a majority of eukaryotic and prokaryotic proteins undergo N-terminal PTMs at some point (Giglione *et al.*, 2015). These include co-translational NME, N-terminal (Nt)-acetylation, Nt-myristoylation, as well as post-translational Nt-methylation and Nt-ubiquitination (Varland *et al.*, 2015; Figure 1.6). Rare PTMs at protein N-termini include Nt-propionylation and Nt-palmitoylation. These modifications are catalysed by a diverse array of enzymes, including peptidases (e.g. methionine aminopeptidases, MetAPs), transferases (e.g. N-terminal acetyltransferases, NATs), and ligases (e.g. E3 Ub ligase). Due to the ubiquity of N-terminal PTMs, proteins with unmodified α-amino groups (i.e. free N-termini) only constitute a small fraction of the total proteome in eukaryotes. In contrast, it is estimated that up to 70 % of the prokaryotic proteome have free N-termini (Ciechanover, 2005).



N-terminal	Enzymes
N-terminal methionine excision	MetAPs
Acetylation	NATs
Myristoylation	NMTs
Methylation	NTMTs
Ubiquitination	Ubiquitin ligase

Figure 1.6 Major co- and post-translational modifications (PTMs) at protein N-terminus (adapted from Varland *et al.*, 2015). MetAPs: methionine aminopeptidases; NATs: N-terminal acetyltransferases; NMTs: N-terminal myristoyltransferses; NTMTs: N-terminal methyltransferases.

N-terminal modifications significantly contribute to protein stability and activity, and their involvement in cellular functions and pathological conditions are starting to be understood in recent years. NME and Nt-acetylation are two widespread PTMs. In eukaryotes, NME is estimated to occur on more than 50 % of all proteins, whereas Nt-acetylation affects 50 % and 80 % of proteins in yeast and humans, respectively (Giglione *et al.*, 2015). In comparison, more than 50 % of prokaryotic proteins undergo NME (catalysed by peptide deformylases and MetAPs) but Nt-acetylation is rare. Recent studies reported that the principal function of these two modifications in eukaryotes is to create a specific protein degradation signal for the N-end rule pathway, forming a major branch called "Ac/N-end rule pathway" (Varshavsky, 2011, Lee *et al.*, 2016). Other functions of Nt-acetylation include intracellular membrane targeting and protein complex formation (Varland *et al.*, 2015).

Nt-myristoylation is the addition with a myristoyl group to an N-terminal G residue, catalysed by N-terminal myristoyltransferases. This modification takes place on 0.5 – 4 % of cellular proteins and is important in cell signalling in cancer or immune responses (Martinez *et al.*, 2008). Protein methylation is a pervasive and functionally versatile PTM. For instance, histone proteins are subject to a range of PTMs including mono-, di- or tri-methylation on a K residue, and mono- or di-methylation on an R residue. These modifications result in the alteration of chromatin structure and thus gene expression (Murn and Shi, 2017). In comparison, Nt-methylation is less well understood. Recent studies reported the involvement of this N-terminal PTM in protein-protein/DNA interactions (reviewed in Varland *et al.*, 2015). Similar to Nt-methylation, Nt-ubiquitination is another emerging PTM, and it is functionally separated from the N-end rule pathway. Nt-ubiquitination has been suggested to perform protein quality control or to regulate protein homeostasis, but currently it is difficult to draw conclusions with limited information. Taken together, all these N-terminal PTMs contribute to a greater proteome complexity, thus presenting analytical challenges to MS-based proteomics.

1.3 High-throughput techniques in N-terminalomics

As briefly covered above, N-terminalomics focuses on the analysis of protein N-termini at the proteome level. A principal objective of N-terminalomics is to reduce the complexity of the proteome. As Beynon (2011) pointed out, standard shotgun proteomics often falls into "over-determinism". Not only are more peptides analysed than strictly necessary to achieve the study aim, but also low-abundance proteins or their component peptides are often overshadowed by the abundant ones in the same sample. Specifically, protein N-termini are

often undetected in standard shotgun proteomics, largely owing to the sheer number and dynamic range of tryptic peptides in a sample (Hartmann and Armengaud, 2014). By selecting highly informative peptides such as protein N-termini, N-terminalomics may improve protein identification and achieve accurate annotation of their N-termini, which can be apparently different from those predicted at gene or transcript level (Rogers and Overall, 2013). With respect to proteolytic processing, this PTM gives rise to new protein fragments with a *neo*-N-terminus (or a *neo*-C-terminus). As a result, N-terminalomics is suitable for analysing proteolytic products as well. Such approaches can be employed to identify *in vivo* proteolytic processing events, or the substrates of a specific protease.

N-terminalomics generally takes advantage of the semi-unique chemistry of the α-amino group to chemically modify protein N-termini for enrichment. Compared to protein N-termini, the C-termini are a lesser target for chemical modifications due to the lower reactivity of carboxyl groups. As a consequence, fewer C-terminalomic techniques have been developed so far (reviewed in Tanco *et al.*, 2015). However, there is an intrinsic issue associated with N-terminalomics: "the lysine problem" (Rogers and Overall, 2013). It describes the similar chemical reactivity between protein N-termini which have α-amino groups and K residues which have ε-amino groups. For instance, the two amino groups only have a marginal difference in their basic strengths expressed as pK_a values: the value of α-amino groups is ~ 9.0, whereas that of ε-amino groups is approximately 10.5 (Dawson *et al.*, 2002). As a result, amine-reactive chemistry generally targets both protein N-termini and K residues simultaneously.

In N-terminalomics, several strategies have been devised to try to circumvent the lysine problem by modifying both protein N-termini and K residues. Alternatively, there are emerging strategies that can discriminate between α- and ε-amino groups. All these strategies comprise the two branches in N-terminalomics: positive or negative selection of N-terminal peptides. The former directly enriches the N-terminal peptides, whereas the latter depletes all the other peptides to select protein N-termini (Figure 1.7A).

1.3.1 Negative selection strategies

One of the first negative selection strategies for N-terminalomics is N-terminal COFRADIC (combined fractional diagonal chromatography), developed by the Gevaert lab at VIB in Belgium (Gevaert *et al.*, 2003). It is a negative selection strategy that combines indiscriminate amine-reactive chemistry with a series of LC steps to separate protein N-termini from other internal peptides (unwanted tryptic peptides) prior to MS analysis (Figure

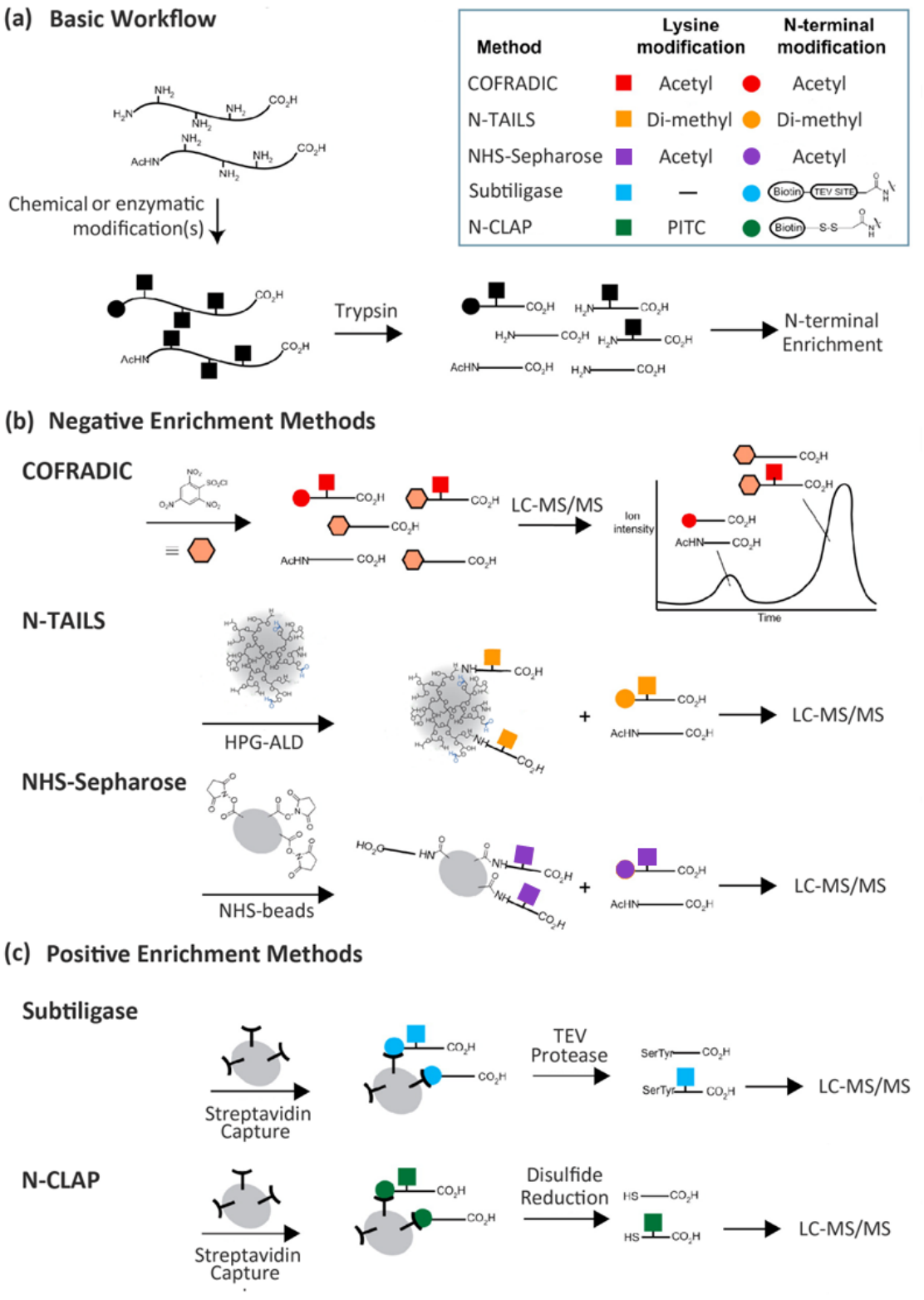


Figure 1.7 Overview of major N-terminalomic techniques (adapted from Agard and Wells, 2009). (**a**) Typical workflow of N-terminalomics. (**b**) Some major negative selection strategies. (**c**) Some major positive selection strategies. PITC: phenyl isothiocyanate; HPG-ALD: hyperbranched polyglycerol-aldehydes; TEV: tobacco etch virus.

1.7B). The latest protocol of N-terminal COFRADIC employs an *N*-hydroxysuccinimide (NHS) ester to acetylate both N-terminal α-amines and K ε-amines, often with isotopic labels for comparative or quantitative analysis. After tryptic digestion of the acetylated proteins, the resulting peptides are subjected to low-pH SCX, which removes the majority of internal peptides. The enriched peptides are then separated by the first run of RP-HPLC, followed by the reaction between *neo*-N-termini and 2,4,6-trinitrobenzenesulfonic acid (TNBS). This reaction confers additional hydrophobicity only to the remaining internal peptides and protein C-termini, and it is followed by the second run of RP-HPLC. Peptides that exhibit a dramatic shift in their RP-HPLC retention time (*RT*) correspond to the internal or C-terminal peptides, which are removed accordingly. Eventually, the original N-terminal peptides are enriched for LC-MS/MS analysis (Staes *et al.*, 2017).

Since its initial publication in 2003, N-terminal COFRADIC has been used for the global analysis of protein N-termini in yeast, fruit flies, mice, and humans. With the maturation of this technology, the number of identified protein N-termini has drastically expended from < 400 to ~ 2,000 (Arnesen *et al.*, 2009, Gawron *et al.*, 2016). The main objectives of these studies are to determine the state of protein Nt-acetylation or alternative translation initiation. The majority of eukaryotic protein N-termini were reported to undergo Nt-acetylation, ranging from 50 – 60 % in yeast to > 80 % in humans (Arnesen *et al.*, 2009). On the other hand, these studies linked about 10 % of the identified protein N-termini in yeast and 20 % in humans to alternative sites for translation initiation (Helsens *et al.*, 2011, Van Damme *et al.*, 2014).

To identify protease substrates or proteolytic processing events, N-terminal COFRADIC requires a sophisticated strategy of isotopic labelling and quantitation methods. Isotopic labelling may be achieved before protein extraction using SILAC or with trideutero(D_3)-labelled acetylation reagents after protein extraction, depending on the sample type and study aim. Nevertheless, this method identified 61 caspase substrates in humans using recombinant caspases (Wejda *et al.*, 2012). Similarly, this technique was extensively employed to study the substrates and cleavage specificity of granzymes in mice and humans (Plasman *et al.*, 2014). Several variants of N-terminal COFRADIC have been proposed to select N-terminal peptides, where TNBS is replaced by other amine-reactive compounds (e.g. Bland *et al.*, 2014a).

Two similar chromatographic approaches have been proposed in recent years, which separate internal peptides from protein N-termini via charge reduction or charge reversal. The charge reduction approach, namely ChaFRADIC (charge-based fractional diagonal

chromatography), involves protease digestion and the acetylation of primary amines on the resulting internal peptides (Venne *et al.*, 2015). Acetylation results in a reduction of peptide charge, which is made use of by SCX chromatography to deplete internal peptides. Recently, the same research group provided a simpler format of this approach, where chromatographic separation was replaced by the use of pipette tips containing SCX beads (Shema *et al.*, 2018). Meanwhile, the charge reversal approach can add up to four negative charges to the internal peptides via disulfo-modification (Lai *et al.*, 2015). This charge reversal allows SCX to enrich for protein N-termini, since the negatively charged internal peptides do not bind to the SCX column.

N-TAILS (N-terminal amine isotopic labeling of substrates) is one of the most widely adopted strategies in N-terminalomics. This is a negative selection strategy developed by the Overall lab at the University of British Columbia (Kleifeld *et al.*, 2010). Similar to N-terminal COFRADIC, the N-TAILS approach also blocks both α- and ε-amino groups before protease digestion, and then depletes internal peptides through their reactive *neo*-N-termini. In detail, the first step of N-TAILS is to block all primary amines by either isotopic dimethylation or isobaric tagging (e.g. iTRAQ). Subsequent protease digestion of the modified proteins gives rise to N-terminal peptides and internal peptides. Each internal peptide contains a reactive *neo*-N-terminus that facilitates its capture by a commercially available polymer (HPG-ALD, 100 kDa) through amine-reactive chemistry. The captured internal peptides can be removed from the N-terminal peptides by ultrafiltration. As a result, the protein N-termini are selected for LC-MS/MS analysis.

N-TAILS was initially developed to identify protease substrates since proteolytic products can be enriched together with protein N-termini. After the negative selection, proteolytic products are identified by quantitative proteomics, which compares the relative abundance of differentially labelled peptides between control and protease-treated samples (Demir *et al.*, 2017). This approach has been successfully applied to reveal both *in vitro* and *in vivo* proteolytic processing events. For instance, matrix metalloproteinases (MMPs) have been extensively investigated by N-TAILS *in vitro*, which identified 146 substrates of MMP-2 and novel substrates of MMP-9 and MMP-10, albeit in lower numbers (Prudova *et al.*, 2010, Schlage *et al.*, 2014). Similarly, N-TAILS identified prothrombin and eight components of the complement system as *in vivo* substrates of macrophage-specific MMP-12, thus linking this protease to inflammation (Bellac *et al.*, 2014). In addition, Klein *et al.* (2015) applied N-TAILS to study immortalised B cells from a patient with a homozygous mutation of paracaspase

MALT1. This study reported HOIL1 protein as a novel substrate of MALT1, and that MALT1 functioned via HOIL1 cleavage to negatively regulate the nuclear factor (NF)-κB signalling.

In addition to proteolytic processing, N-TAILS has also been heavily involved in the global analysis of protein N-termini and their PTM states in various biological systems. For instance, N-TAILS successfully identified 1,400 protein N-termini from human erythrocytes, with 68 % of them derived from alternative sites for translation initiation or proteolytic processing (Lange *et al.*, 2014b). Dental pulp and plants represent additional biological systems that have also been subject to N-TAILS analysis (reviewed in Demir *et al.*, 2017). Due to the relatively wide use of N-TAILS, the applications of this method are certainly not exhaustively covered here.

By analogy with N-TAILS, a negative selection strategy can be devised by combining other amine-reactive chemistry (to block all primary amines) with the capture of internal peptides. The PTAG (phospho-tagging) strategy exploits the AP of phosphopeptides to deplete internal peptide and thus enrich for protein N-termini (Mommen *et al.*, 2012). This strategy consists of protein dimethylation (to block all primary amines), protease digestion, phosphorylation of internal peptides (through amine-reactive chemistry), and TiO_2 affinity chromatography (to remove phosphopeptides). Consequently, the dimethylated protein N-termini are negatively selected.

Prior to the advent of N-TAILS and PTAG, the Beynon group at the University of Liverpool developed a negative selection strategy based on a similar concept (McDonald *et al.*, 2005). In detail, this approach (here referred to as “NHS-Sepharose”) involves the blocking of both protein N-termini and K residues by (isotopic) acetylation. Here the use of isotopes affords discrimination between endogenous and chemical acetylation. The acetylated proteins are then digested with a protease. Initially, the resulting internal peptides are chemically tagged via the reaction between *neo*-N-termini and an amine-reactive biotin tag. The biotin tag mediates the removal of internal peptides by immobilised biotin-binding proteins such as avidin or streptavidin. This procedure has since been simplified: the internal peptides are now directly depleted using NHS-activated Sepharose, which reacts with primary amines to form a stable amide bond (McDonald and Beynon, 2006). Despite its early invention, this technique has not been widely adopted in N-terminalomics to reduce sample complexity. Therefore, its full potential remains largely unknown.

Binding of biotin to avidin or streptavidin is one of the strongest non-covalent interactions between a protein and its ligand (Green, 1975). Biotin is a water-soluble vitamin, and it

serves as a cofactor for mammalian carboxylases that are critical to metabolic processes including fatty acid synthesis (reviewed by Said, 2012). Avidin is a tetrameric biotin-binding protein (67 kDa) from chicken egg-white. Each subunit (128 amino acid residues) of chicken avidin binds to a biotin molecule with an exceptionally high affinity, i.e. $K_d \approx 10^{-15}$ M (Marttila *et al.*, 2000). In contrast, streptavidin (56 kDa) is a bacterial avidin homologue from *Streptomyces avidinii*. Streptavidin and avidin share modest sequence homology (38 %) but highly similar protein structures (Dundas *et al.*, 2013). Compared to chicken avidin, streptavidin is not glycosylated and more acidic, but with a similar binding affinity ($K_d \approx 10^{-14}$ M).

Both avidin and streptavidin maintain their stability and functions over a wide range of pH and temperature, and tolerate genetic mutations (Gonzalez *et al.*, 1999, Dundas *et al.*, 2013). Therefore, biotin-(strept)avidin interaction has been heavily utilised in molecular biology and biotechnology. At present, there are a series of avidin and streptavidin variants that are commercially available. These include NeutrAvidin, a deglycosylated avidin that exhibits reduced nonspecific binding (Marttila *et al.*, 2000), and monomeric avidin, which allows gentle elution of the bound biotin molecules due to a much lower affinity (Henrikson *et al.*, 1979).

1.3.2 Positive selection strategies

In contrast to negative selection, protein N-termini can in principle be directly enriched using positive selection strategies. Generally, such strategies exploit chemical or enzymatic reactions that discriminate between α- and ε-amino groups. Due to the extent of endogenous Nt-acetylation of eukaryotic proteins, such strategies are more suitable for studying proteolytic processing. Positive selection strategies are represented by the Subtiligase approach, which was developed by the Wells lab at the University of California at San Francisco (Mahrus *et al.*, 2008). Subtiligase is an engineered enzyme highly specific for α-amino groups but not the ε-ones. The absolute specificity has been exploited by the Wells lab to transfer a biotin tag from synthetic peptide ester substrates to free (*neo*-)N-termini (Figure 1.7C). The peptide ester itself consists of four components: an ester linkage (for subtiligase-catalysed ligation to α-amino groups), a biotin tag (for AP), a tobacco etch virus (TEV) protease cleavage sequence (for releasing the captured peptides), and an α-aminobutyric acid (Abu) mass tag (for confident identification by MS).

As described in the latest protocol, slight variations exist between *in vivo* and *in vitro* Subtiligase experiments (Wiita *et al.*, 2014). Nevertheless, the approach at its core involves

the initial enzymatic biotinylation of protein N-termini and proteolytic products at their *neo*-N-termini. After the removal of excess biotin reagents (i.e. the synthetic peptide ester) through protein precipitation, the biotinylated protein N-termini/proteolytic products are captured by immobilised NeutrAvidin through biotin-avidin interaction. On-bead tryptic digestion is then performed on the captured proteins to release unwanted internal peptides. The captured peptides are finally eluted through TEV protease digestion for LC-MS/MS analysis.

Identification of caspase substrates has greatly benefited from this strategy, as the first Subtiligase experiment reported > 1,000 proteins caspase substrates in apoptotic cells (Mahrus *et al.*, 2008). This study initiated the compilation of specific substrates of different caspases using this strategy (Agard *et al.*, 2010, Agard *et al.*, 2012). Subtiligase has also been employed to profile protein N-termini in human blood (Wildes and Wells, 2010). This study identified 772 N-termini in human blood, of which 28 % were derived from cleavage by aminopeptidases.

N-CLAP (N-terminalomics by chemical labeling of the α-amine of proteins) and *O*-methylisourea are two additional positive selection strategies. N-CLAP employs Edman degradation chemistry that specifically cleaves off the N-terminal amino acid residue from a protein (Xu *et al.*, 2009). An N-CLAP experiment starts with the modification of both α- and ε-amino groups with phenyl isothiocyanate (PITC). The PITC-modified N-terminal residues are then removed from the rest of the proteins following the addition of trifluoroacetic acid (TFA), which triggers cleavage of the peptide bond between PITC-modified N-termini and their adjacent residues. The resulting *neo*-N-termini of the remaining proteins can be exploited for biotin tagging and AP via biotin-avidin interaction (Xu and Jaffrey, 2010).

Similarly, *O*-methylisourea exhibits selective reactivity towards the ε-amino groups of K residues under defined conditions. The Salvesen group (2011) at Cornell University thus developed a positive selection approach based on selective guanidination with *O*-methylisourea. This approach blocks K ε-amino groups to a large extent, leaving the free protein N-termini available for biotin tagging and AP (after tryptic digestion). Furthermore, selective guanidination was incorporated into another positive selection approach taking advantage of thiol-specific AP (Kim *et al.*, 2013). However, the use of *O*-methylisourea is not ideal as it incompletely modifies proteins (Cohen, 1968). In addition, this compound is not absolutely specific for ε-amino groups. Guanidination of N-terminal α-amino groups has been observed in proteins with N-terminal G or other residues due to reduced steric protection (Beardsley *et al.*, 2000).

Although the last decade has seen many technological breakthroughs in N-terminalomics, there is clearly ample room for innovation, especially in the positive selection branch. It owes primarily to the difficulty in developing chemical reactions that can discriminate between α- and ε-amino groups. In view of the initial reports from Sonomura *et al.* (2009a), selective transamination of protein N-termini seemed to be a promising reaction for the development of a positive selection approach. The following section will provide an in-depth description of this chemical reaction.

1.4 Transamination of protein N-termini

In nature, transamination refers to the transfer of an amino group from amino acids to α-ketoacids, most notably to α-ketoglutarate. This reaction is catalysed by a large variety of aminotransferases that contain a prosthetic group called pyridoxal-5'-phosphate (PLP). PLP is also required for other enzymatic reactions on amino acids, including decarboxylation, racemisation, and β-elimination (Toney, 2005). Enzymatic transamination is an intrinsic component of the metabolism of amino acids (Berg *et al.*, 2012). On the other hand, the development of chemical transamination of protein N-termini was primarily contributed by Dixon (1964), who developed this reaction based upon earlier work on free amino acids (Mix and Wilcke, 1960).

1.4.1 Selective transamination of protein N-termini and other transamination routes

As previously described, selective modification of the N-terminal α-amino group is hampered by the presence of ε-amino groups on the side chains of K residues, which are 25-fold more abundant on average (the lysine problem; Mahrus *et al.*, 2008). From a historical perspective, there have been several chemical reactions that target the α-amino group at the N-terminus. The selectivity of such modifications was achieved (to a limited extent) through the careful control of reaction pH, on the basis that α- and ε-amino groups have slightly different pK_a values (see section 1.3; Reid, 1951, Baker *et al.*, 2006). On the other hand, specific types of N-terminal amino acid residues have been successfully targeted via their functional groups on the side chain: proteins with an N-terminal cysteine (C) were modified with thioesters, whereas N-terminal S and T residues were converted to reactive aldehydes using periodate (Dixon and Weitkamp, 1962, Dawson *et al.*, 1994). It was possible to further convert the reactive aldehydes to G residues (Fields and Dixon, 1968).

In spite of these individual cases of success, it is desirable to devise a general strategy that targets virtually all types of amino acid residues at protein N-termini. This combination of broad spectrum with high specificity may be achieved by an N-terminal modification if it directly involves the neighbouring peptide bond, which ε-amino groups lack. With this in mind, a selective transamination reaction was developed to modify protein N-termini with three essential components: an amino group acceptor, a heavy metal cation, and a high concentration of base (reviewed in Dixon and Fields, 1972). Although there is considerable flexibility in the choice of reagents, the preferred reaction system uses glyoxylate as the amino acceptor, copper(II) as the metal ion (e.g. using $CuSO_4$), and pyridine as the base (Dixon, 1984). Substitution of pyridine with sodium acetate results in a milder reaction condition, which in principle allows the reaction to proceed without protein denaturation.

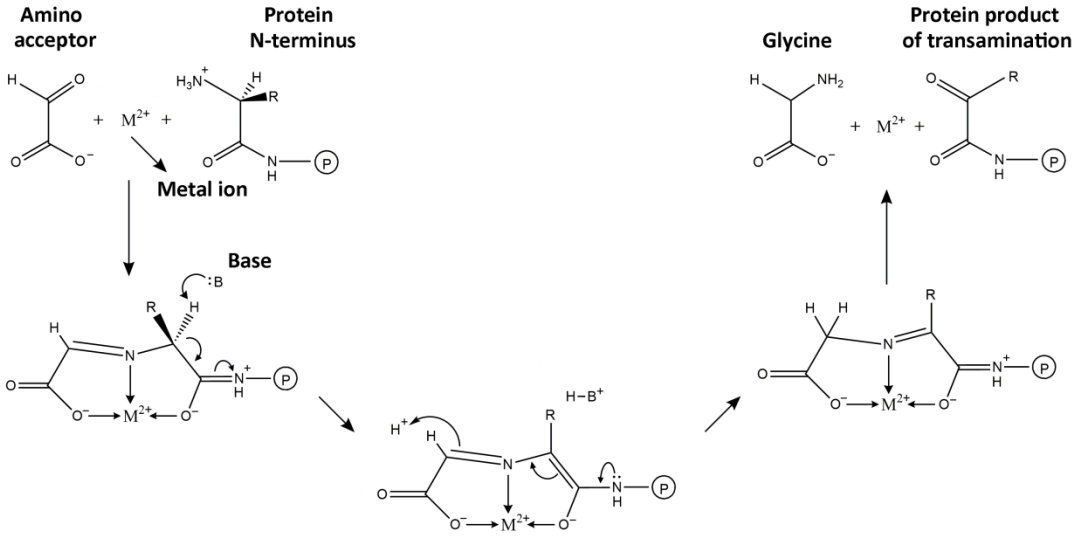


Figure 1.8 Schematic diagram of the selective transamination of protein/peptide N-termini (adapted from Dixon, 1984). The negatively charged oxygen of an amino acceptor (e.g. glyoxylate) is first juxtaposed with the carbonyl oxygen of the first amino acid residue of a protein or peptide Ⓟ via a divalent metal cation. The presence of a base (B) then allows the formation of an imine which, following a sequence of proton transfer events, is hydrolysed thereby converting the α-amine to a reactive carbonyl group.

As depicted in Figure 1.8, divalent metal ions catalyse the reaction by first chelating both the carbonyl group in the peptide bond and the carboxyl group of glyoxylate, which results in imine formation between the N-terminal α-amino group and glyoxylate. Following a series of proton transfer events (mediated by the base and the peptide bond), the imine is hydrolysed to complete the conversion of the α-amine to a reactive carbonyl group. This newly formed carbonyl group is amenable to further chemical derivatisation.

Dixon's pioneering work inspired other researchers to develop transamination of protein N-termini from different perspectives. For the sake of clarity, only the Dixon's method will be referred to as "selective transamination" in this thesis. The 21st century has seen the resurgence in the development and use of protein Nt-transamination. For instance, the Francis group at the University of California at Berkeley evaluated the potential of two molecules for transamination: PLP and *N*-methylpyridinium-4-carboxaldehyde (Rapoport's salt, RS). As mentioned before, PLP is a natural cofactor for a great many enzymes that catalyse the transformation of amino acids in metabolism. Gilmore *et al.* (2006) reported that, under optimised conditions (e.g. pH 6.5, 37 °C, 20 h), treatment with PLP alone led to Nt-transamination on a variety of proteins with up to 80 % reaction yield. Using a library-based screening strategy, an N-terminal AKT motif (A = alanine) was later determined to improve the yield of PLP-mediated transamination (Witus *et al.*, 2010).

The same screening strategy was also implicated in the development of RS-mediated Nt-transamination. Witus *et al.* (2013) reported that synthetic peptides with an N-terminal glutamate (E) residue were generally amenable to this reaction under mild conditions (pH 6.5, 1 h). In the peptide library, over 80 % of the peptides with an N-terminal EE motif were transaminated. By elevating the reaction temperature to 37 °C, the N-termini of human IgG1 antibody were successfully transaminated, especially on the heavy chain. It was also shown that the antigen binding property of IgG1 antibody was not disrupted by transamination and subsequent carbonyl modification.

Although both the PLP and RS routes have been successfully employed to modify protein N-termini, the authors did report their respective limitations. For instance, the PLP route typically takes up to 24 h to complete. In addition, elevating the temperature does shorten the reaction time but at the cost of protein denaturation. In addition, the reaction yield is highly variable and depends on the identity of the N-terminal residue, and the desired product is often mixed with unwanted PLP adducts (Rosen and Francis, 2017). With respect to the RS route, the reaction scope is largely limited to proteins with an N-terminal E residue. Furthermore, a proline (P) residue at the second position adversely affects the efficiency of transamination (Witus *et al.*, 2013).

1.4.2 Applications of Nt-transamination

An outstanding feature of protein Nt-transamination is its versatility in potential applications. This reaction selectively destroys the N-terminal α-amino group, but introduces a reactive carbonyl group instead. Generally the carbonyl group is not naturally present in proteins, thus it can be exploited to further introduce desired functionalities. Owing to this flexibility, selective transamination has been utilised in a wide range of biochemical studies. Initially, this reaction was applied by Dixon and Moret (1964) to specifically remove the N-terminal residue from a protein of interest. In detail, the novel carbonyl group (introduced by transamination) was further derivatised with a dual-functional nucleophile (e.g. *O*-phenylenediamine). Subsequently, the nucleophile would attack the neighbouring peptide bond to release the N-terminal residue. Scission of the N-terminal residue allowed the researchers to determine the relevance of the N-terminal residue in the function of pig corticotropin, bacterial cytochrome C, and human serum albumin (Dixon and Moret, 1964, Dixon and Moret, 1965, Sarkar *et al.*, 1978).

In contrast to the early work that focuses on the biochemistry of protein N-termini, recent studies take advantage of the unique carbonyl group for site-specific protein labelling (i.e. protein bioconjugation). Such studies typically employ alkoxyamine compounds (e.g. *O*-benzylhydroxylamine, $BnONH_2$) to form an oxime bond with the carbonyl group. For instance, $BnONH_2$ has been involved in the development of both the PLP and RS routes of Nt-transamination. In these studies, $BnONH_2$ was primarily used to validate the transamination products and to determine the reaction yields. Other alkoxyamine derivatives facilitated the bioconjugation of polyethylene glycol (PEG) polymers and fluorescent probes to protein N-termini (reviewed in Rosen and Francis, 2017). The application of Nt-transamination has also been extended to MS-based peptide sequencing. In such studies, selective transamination is combined with further carbonyl modifications to negatively select either the N- or C-terminal peptide of target proteins (Figure 1.9).

To enrich for N-terminal peptides, selective transamination was performed to block the N-terminal α-amino group prior to further modification with an alkoxyamine compound (Sonomura *et al.*, 2009a). After digestion with a specific protease (LysN), the internal peptides all started with a K residue. These internal peptides were captured by amine-reactive groups that were immobilised on glass. The N-terminal peptides did not contain any reactive primary amine and thus resisted capture. On the other hand, selection of C-terminal peptides required LysC protease digestion and peptide transamination, which rendered the C-terminal peptides resistant to removal by the same amine-reactive glass

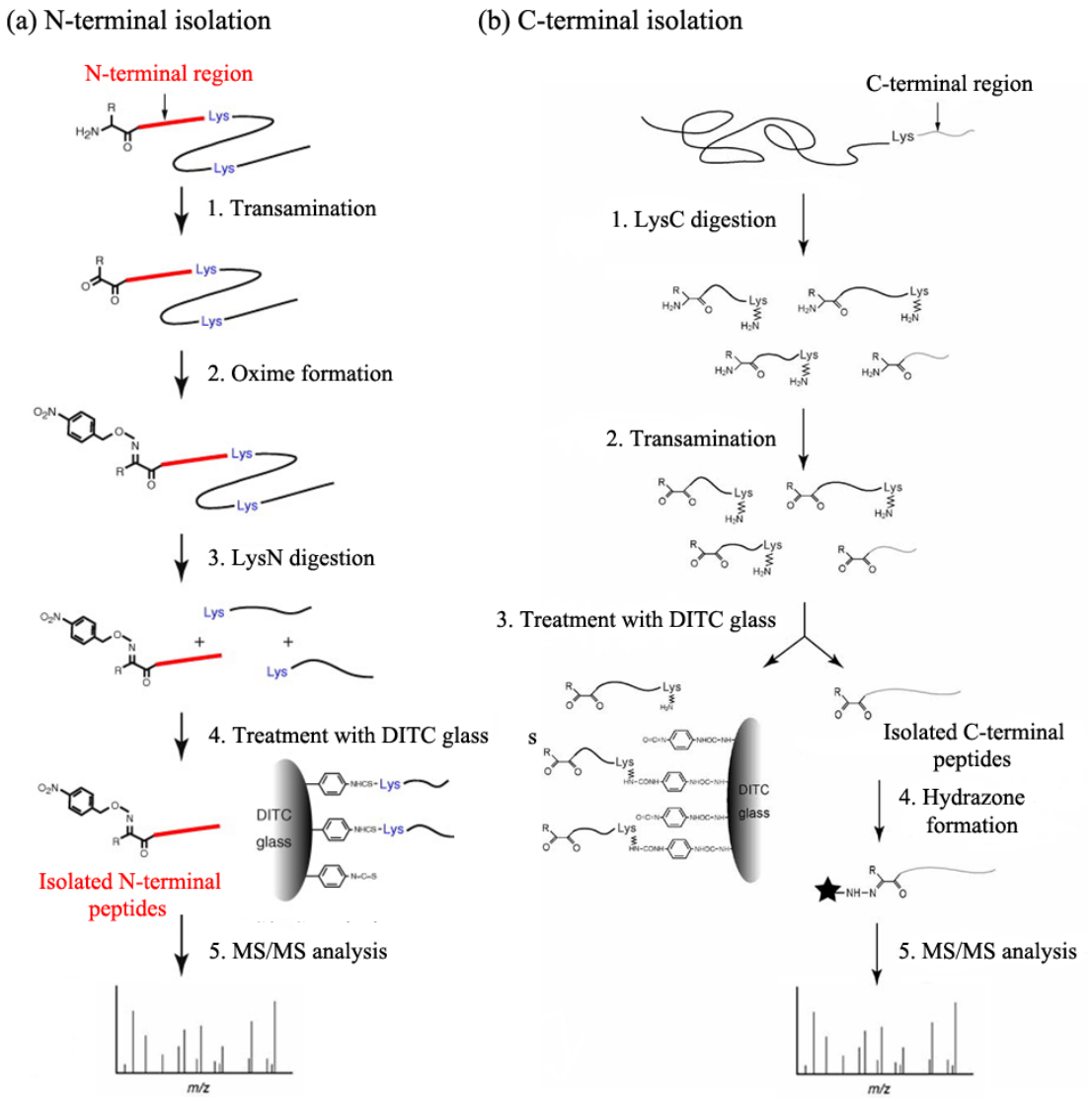


Figure 1.9 Schematic diagram of the negative selection of protein N-termini (**a**) or C-termini (**b**) by selective transamination (adapted from Sonomura *et al.*, 2009a, Sonomura *et al.*, 2009b). DITC: *p*-phenylenediisothiocyanate.

(Sonomura *et al.*, 2009b). Subsequently, the C-terminal peptides were blocked through oxime formation. The selected N- or C-terminal peptides were sequenced by MS in the end.

In addition to oxime formation, the unique carbonyl group is also amenable to hydrazone formation, reductive amination, and Pictet-Spengler reaction. For instance, Sonomura *et al.* (2011) provided an upgraded transamination strategy to select endogenously modified protein N-termini. This method takes advantage of an immobilised hydrazine derivative to directly scavenge the transaminated internal peptides through hydrazone bond formation. Naturally blocked protein N-termini do not undergo selective transamination due to the absence of a reactive α-amino group, and are thus selected for MS analysis.

On the other hand, Sasaki *et al.* (2008) reported Nt-biotinylation on horse heart myoglobin through the PLP route of transamination and the Pictet-Spengler reaction. In this study, the newly introduced carbonyl group was modified with a biotin-derivatised tryptamine under mild conditions (pH 6.5, 37 °C). The carbonyl group underwent cyclic condensation to form a ring structure that contained a stable C-C bond, and this reaction did not cause any major change in protein structure. The Bertozzi group at Stanford University further improved the Pictet-Spengler reaction through rational design of tryptamine analogues, leading to a 50 % increase in reduction efficiency (Agarwal *et al.*, 2013). Similar to the study by Sasaki *et al.* (2008), the improved reaction also enabled Nt-biotinylation of horse heart myoglobin as a proof-of-concept. In addition, a fluorescent dye was selectively attached to the N-terminal carbonyl group of modified human IgG1 antibody.

1.4.3 Alternative amine-reactive chemistry

In addition to the transamination of N-termini, several chemical modifications have also been proposed to target primary amines including the α-amino groups at protein N-termini. As illustrated in Figure 1.10, *N*-succinimidyl *S*-acetylthioacetate (SATA) and citraconic anhydride (CA) represent two types of amine-reactive compounds that do not discriminate between α- and ε-amino groups. SATA is a dual-functional compound that reacts with primary amines through the NHS group and introduces a thiol group in a protected form. As described in section 1.3.2, Kim *et al.* (2013) employed SATA to introduce a protected thiol group at protein N-termini as all the ε-amino groups of K residues had been blocked by guanidination with *O*-methylisourea. Following protease digestion, the protected thiol group was liberated with hydroxylamine and then reacted with thiopropyl Sepharose for positive selection of N-terminal peptides.

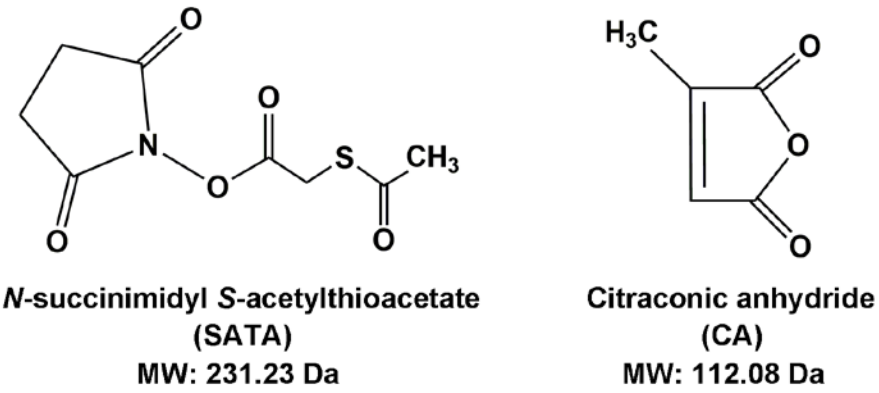


Figure 1.10 The chemical structures of *N*-succinimidyl *S*-acetylthioacetate (SATA, left) and citraconic anhydride (CA, right; modified from Hermanson, 1996).

On the other hand, CA reacts with primary amines in a reversible manner, forming an amide bond under alkaline conditions (pH 7 – 9) that readily dissociates at pH 3 – 4 (Dixon and Perham, 1968). It was employed by the Zhou and Vogelstein groups in the development of a protein bioconjugation technique called PRINT (PRotect, INcise, Tag; Sur *et al.*, 2015). This technique requires a rational design of recombinant proteins, which will contain an N-terminal affinity tag (e.g. His-tag) followed by a TEV protease cleavage site. After expression and purification, the recombinant proteins are modified with CA to temporarily block both α- and ε-amino groups. Digestion with TEV protease exposes a reactive α-amino group, which can be modified through amine-reactive chemistry. For instance, recombinant tumor necrosis factor-α (TFT-α) protein was subject to PRINT for Nt-PEGylation. Without alterations in biological activity, this modification conferred higher stability and reduced toxicity to the recombinant TFT-α.

In contrast to the above compounds, several reagents exhibit high selectivity for the N-terminal α-amino group over the ε-amino groups on K side chains. These include ketenes, *N*-succinimidyloxycarbonylmethyl tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP), *O*-aminophenols, and 2-pyridinecarboxaldehyde (2PCA). Chan *et al.* (2012) demonstrated the feasibility of using an alkyne-functionalised ketene to modify the N-termini of both synthetic peptides and proteins. This reaction was shown to have a broad substrate scope, which is similar to selective transamination. All 20 natural amino acids except histidine (H) could be modified with modest to excellent specificity when positioned at the N-terminus of a peptide. In addition, this N-terminal modification introduced a novel alkyne group, which could facilitate further functionalisation via click chemistry.

TMPP is another commonly used reagent for labelling protein N-termini. Not only specific for the α-amino groups, TMPP modification is also beneficial to peptide identification by MS (Huang *et al.*, 1997). Hence, this compound was used in a gel-based approach to assist in robust identification of protein N-termini (Deng *et al.*, 2015). However, this method is relatively labour-intensive and low throughput. TMPP has been further combined with selective transamination or COFRADIC for negative selection and high-throughput identification of N-terminal peptides (Sonomura *et al.*, 2011, Bland *et al.*, 2014b). However, the potential of TMPP has yet to be fully exploited due to a lack of strategies for AP. A TMPP-specific antibody was recently reported (Bland *et al.*, 2014a), but it has neither been thoroughly scrutinised nor is it commercially available to our knowledge.

The Francis group reported both oxidative coupling with *O*-aminophenols and 2PCA condensation as N-terminal labelling approaches. The oxidative coupling approach involves

the oxidation of *O*-aminophenols by $K_3Fe(CN)_6$ and *in situ* reaction of the oxidised intermediates with protein α-amino groups (Obermeyer *et al.*, 2014). Meanwhile, 2PCA undergoes cyclic condensation with the N-terminal α-amino groups to form an imidazolidinone product (MacDonald *et al.*, 2015). However, it was reported that the oxidative coupling approach favoured the N-terminal P residue, whereas the imidazolidinone products from 2PCA condensation were relatively unstable at 37 °C (reviewed in Rosen and Francis, 2017).

Although some of the above approaches are efficient and specific towards protein N-termini, selective transamination presents several advantages: first, all the reagents in this method are easily accessible; second, it is versatile in introducing further functionalities; finally, this reaction proceeds efficiently. Although this reaction has been employed to negatively select the N-terminal peptide of individual proteins, its potential in proteomic research has yet to be fully realised. The present work attempts to develop a positive selection strategy based on selective transamination. In N-terminalomics, one crucial element of method development is choosing a suitable biological system: small-scale experiments need to be scaled up on complex protein mixtures. Proteins extracted from Jurkat T-lymphocytes represent such suitable biological system. For the sake of completeness, the following section is devoted to introducing Jurkat T-lymphocytes.

1.5 Jurkat T-lymphocytes as a model system

1.5.1 T-cell biology and the Jurkat model system

Jurkat T-lymphocytes (T-cells, Clone E6-1) are a cancerous cell line, established from a patient with acute T-cell leukaemia in the 1970s (Figure 1.11; Schneider *et al.*, 1977). Initially, Jurkat T-cells gained interest from immunologists for their inducible production of a large quantity of interleukin-2 (IL-2; Gillis and Watson, 1980). There had been a huge demand for IL-2 to maintain T-cell cultures in laboratory, owing to the ability of this cytokine to support T-cell survival and proliferation (Abbas *et al.*, 2018). Jurkat T-cells soon became a major source of IL-2 as well as a valuable model system for T-cell research.

Under physiological conditions, T-cells are responsible for cell-mediated immunity. Together with the humoral immunity mediated by B-lymphocytes (B-cells), cell-mediated immunity constitutes the adaptive immune system, a major part of immune defences in vertebrates. The principal functions of such immune systems are to protect the host against foreign antigens or cancer cells, to establish tolerance to the host's own antigens, and to develop

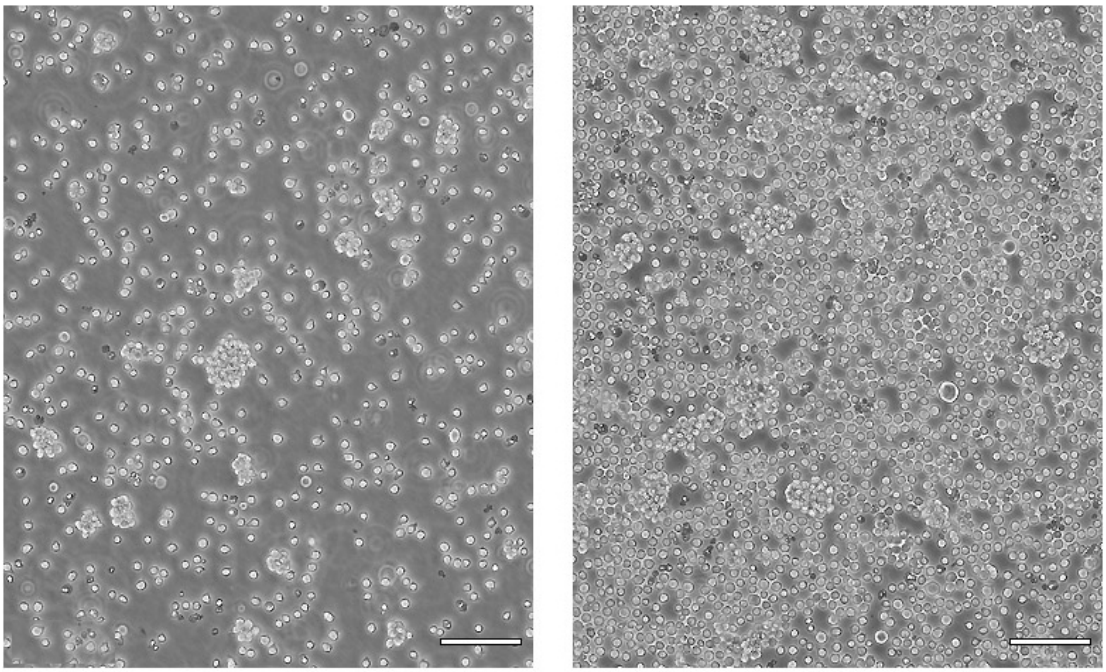


Figure 1.11 Reference images of Jurkat T-cells (source: ATCC, http://www.atcc.org/products/all/TIB-152). Left: low-density; right: high-density; scale bar = 100 µm.

immunologic memory. In contrast to innate immunity that serves as a first line of defence, adaptive immunity is characterised by higher potency, antigen specificity, structural diversity, and immunologic memory. It was acquired later in evolution and is a hallmark of vertebrates (Liongue *et al.*, 2011).

T-cells, such as Jurkat T-lymphocytes, bind to their cognate antigens via a T-cell receptor (TCR). Each T-cell expresses a characteristic TCR that recognises a single antigen. Co-receptors are also present on the surface of T-cells. For instance, the CD3 (cluster of differentiation 3) co-receptor is non-covalently associated with TCR, forming the TCR-CD3 complex (Figure 1.12). T-cells respond to antigens that are displayed on the surface of antigen-presenting cells (APCs), rather than those existing in soluble, cell-free form. Upon entry to the host, antigens may be directly captured by APCs. These cells employ specialised cell surface proteins, the major histocompatibility complex (MHC), to present small peptides derived from the captured antigens to the correct T-cells. There are two classes of MHC molecules: Class I is responsible for presenting intracellular antigens to T cells that express the CD8 co-receptor, whereas Class II presents extracellular ones to T cells expressing the CD4 co-receptor (Abbas *et al.*, 2018).

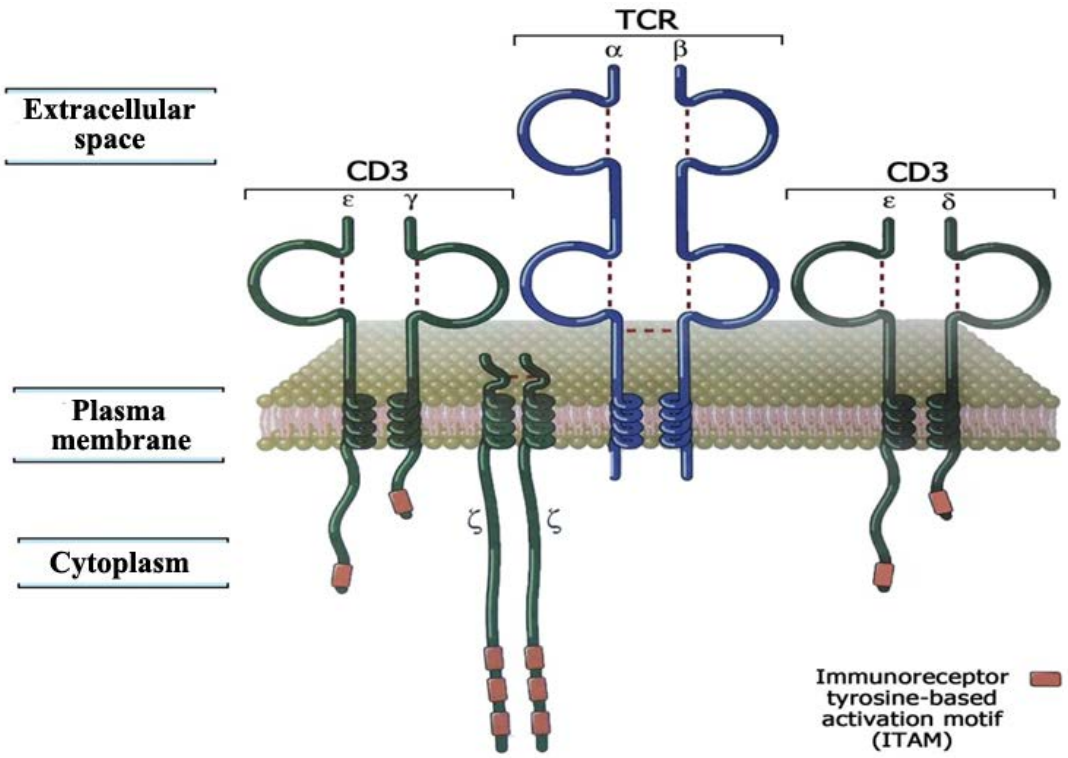


Figure 1.12 Composition of the T-cell receptor (TCR)-CD3 complex, which consists of the TCR, two heterodimers (γε and δε) and a ζζ homodimer of CD3. All cytosolic domains of CD3 proteins contain immunoreceptor tyrosine-based activation motifs (ITAMs). ITAMs become phosphorylated upon TCR ligation and trigger downstream signalling (Abbas *et al.*, 2015). CD3: cluster of differentiation 3.

In secondary lymphoid organs (e.g. lymph nodes and spleen), naïve T-cells browse peptide-MHC complexes on the surface of APCs and become activated after having encountered the correct complex. One of the major contributions from Jurkat T-cells is the two-signal hypothesis of T-cell activation. Weiss *et al.* (1984) determined that neither a monoclonal antibody (i.e. OKT3) against the CD3 co-receptor within the TCR-CD3 complex nor a phorbol ester alone could induce Jurkat T-cells to express a high level of IL-2, but the combination of both boosted IL-2 production. As a result, a simplified model of T-cell activation was put forward consisting of two signals. The first signal (signal 1), derived from the ligation of the TCR-CD3 complex, and the second (signal 2) from phorbol esters. After nearly 35 years of research, the two-signal hypothesis has been refined now: signal 1 is indeed from the ligation of the TCR with a peptide-MHC complex, which is facilitated by the CD4 or CD8 co-receptor; signal 2 comes from the interaction between a T-cell costimulatory receptor (e.g. CD28) and the B7 proteins on APCs (Abraham and Weiss, 2004). These two signals work in synergy to fully activate T-cells, to maintain the survival of activated T-cells, and to promote their proliferation (i.e. clonal expansion) and differentiation (Abbas *et al.*, 2018).

1.5.2 Jurkat proteomics

Since its establishment some forty years ago, Jurkat cell line has been extensively employed in the studies on T-cell signalling. Into the post-genomic era, the use of Jurkat T-cells has been extended to the field of proteomics. The proteomic analysis by Kang *et al.* (2005) represents one of the early attempts to profile global protein expression in Jurkat T-cells. In this study, two LC separations (SCX and RP-HPLC) were online coupled to MS/MS analysis, leading to the identification of 681 proteins. With technological improvements, Geiger *et al.* (2012) identified nearly 8,000 proteins in Jurkat T-cells using a combination of off-line SCX and online RP-HPLC, followed by MS/MS analysis with an LTQ-Orbitrap instrument. Given the importance of protein phosphorylation in T-cell signalling, the global analysis of PTMs is heavily emphasised in Jurkat proteomics. For instance, a time-course analysis of T-cell activation identified more than 5,500 phosphopeptides from 2,008 Jurkat proteins (Nguyen *et al.*, 2016). This study also highlighted the importance of phospholipase C-gamma1 (PLCγ1) in T-cell signalling by comparing the regulated phosphorylation events between the wild-type and a PLCγ1 mutant of Jurka T-cells.

With respect to N-terminalomics, Jurkat T-cells have been involved in the development of both positive and negative selection methods, including N-terminal COFRADIC, Subtiligase, and N-CLAP. In the initial publication of N-terminal COFRADIC (Van Damme *et al.*, 2005), the Gevaert lab reported the use of this technique to directly identify > 58 caspase cleavage sites in apoptotic Jurkat cells. Similarly, the Wells lab demonstrated the utility of Subtiligase by identifying nearly 300 putative caspase substrates in apoptotic Jurkat cells (Mahrus *et al.*, 2008). Additionally, Jurkat proteins were employed in the development of N-CLAP as a model system to study NME, signal peptide removal, and caspase cleavage during apoptosis (Xu *et al.*, 2009). In summary, Jurkat T-cells have already become the backbone of the research on cell-mediated immunity, reflected by the > 17,000 references for “Jurkat” (source: PubMed, 12 Jan. 2018). It is expected that Jurkat T-cells will continue to provide valuable insights into T-cell biology. They will also continue to serve as a valuable reagent and benchmark for testing new proteomic strategies.

1.6 Aims of the present work

The present study aims to employ a range of synthetic peptides, individual proteins, and protein extracts from Jurkat T-cells as model systems to critically evaluate and optimise an existing negative selection approach (NHS-Sepharose) in N-terminalomics. This work also attempts to explore the potential of selective transamination as a positive selection approach. Furthermore, the present work also aims to employ this reaction to improve proteome coverage in standard shotgun proteomics. A detailed break-down of the said aims is shown below:

- To critically evaluate the utility of the NHS-Sepharose approach in selecting the N-terminal peptide of individual proteins for LC-MS/MS analysis.
- To optimise the NHS-Sepharose approach by systematically refining the experiment parameters.
- To evaluate the performance of SATA and CA modifications as amine-reactive chemistries to complement the NHS-Sepharose approach.
- To develop selective transamination as an N-terminalomic approach for tagging the N-termini of synthetic peptides, individual proteins, and Jurkat proteins.
- To evaluate the performance of this approach to positively select protein N-termini for LC-MS/MS analysis.
- To investigate whether selective transamination can be used to improve proteome coverage.

Chapter 2. Materials and Methods

2.1 Materials

The Jurkat T-lymphocyte cell line (ATCC TIB-152™; Schneider *et al.*, 1977) was provided by Dr Hebin Liu and Dr Rong Rong (XJTLU, China). Gibco™ RPMI 1640 medium, PBS (pH 7.4, for cell-culture), and TAE buffer (Tris-acetate-EDTA, 50X) were from Thermo Fisher Scientific, whereas foetal bovine serum (FBS) was purchased from Bovogen Biologicals. Penicillin-Streptomycin was from Sigma-Aldrich, whereas cell culture flasks (T-25 and T-75) and Costar® 96-well plate were obtained from Corning. DNase I, 100 bp DNA ladder and DNA gel loading dye (6X) were from New England Biolabs, whereas GelRed® nucleic acid gel stain was purchased from Biotium.

EZ-Link™ hydrazide-biotin, alkoxyamine-PEG_4-biotin, alkoxyamine-PEG_4-SS-PEG_4-biotin, Pierce™ sulfo-NHS (*N*-hydroxysuccinimide) acetate, NHS-activated magnetic beads, *N*-succinimidyl *S*-acetylthioacetate (SATA), citraconic anhydride (CA), Zeba™ desalting columns (7K MWCO), NeutrAvidin agarose, and monomeric avidin agarose were all purchased from Thermo Fisher Scientific. NHS-activated Sepharose™ 4 Fast Flow was obtained from GE Healthcare. Pyridine and copper(II) sulfate ($CuSO_4$) were products of Alfa Aesar. DL-Dithiothreitol (DTT), iodoacetamide, guanidine hydrochloride (HCl), 2,4-dinitrophenylhydrazine (DNPH), and urea were acquired from Amresco. Amicon® ultrafiltration units (3K MWCO) and ZipTip® C18 pipette tips were from Merck Millipore, whereas trypsin and endoproteinase Glu-C were obtained from Promega.

Biotin and dinitrophenol rabbit monoclonal antibodies were purchased from Cell Signaling Technology, whereas IRDye® 680LT donkey anti-rabbit IgG antibody was from LI-COR. NuPAGE™ Bis-Tris precast gels (10 %, 12-well), antioxidant, SeeBlue Plus2 pre-stained standard (3 – 198 kDa), and the Colloidal Blue staining kit were products of Thermo Fisher Scientific. 5X sodium dodecyl sulfate (SDS) sample buffer, protein molecular weight (MW) marker (14.4 – 116 kDa), and phenylmethanesulfonyl fluoride (PMSF, 100 mM) were from Beyotime, and nitrocellulose membrane was purchased from Pall Life Sciences.

Synthetic human adrenocorticotropic hormone (ACTH, amino acid sequence: SYSMEHFRWG) and rat renin substrate (DRVYIHPFHLLYYS) were provided by Chinese Peptide Company (CPC). Bovine serum albumin (BSA), chicken egg-white lysozyme (lysozyme C), QuantiPro™ bicinchoninic acid (BCA) assay kit, Tris(2-aminoethyl)amine (polymer-bound), *O*-benzylhydroxylamine ($BnONH_2$), and all other chemicals were obtained from Sigma-Aldrich.

2.2 Cell culture

Cryopreserved Jurkat T-lymphocytes (5 × 10^6 cells/ml) were quickly thawed in a water bath at 37 °C, prior to addition of 5 ml of RPMI 1640 medium containing 5 % (v/v) FBS and 1 % (v/v) Penicillin-Streptomycin (10,000 units/ml), and then collection by centrifugation at 200 x *g* for 3 minutes (min). After the removal of the supernatant, the pellet was resuspended in 5 ml of the complete medium, transferred to a T-25 cell culture flask and grown in a 37 °C incubator with 5 % CO_2 for two days. Cells were then sub-cultured by adding 10 ml of fresh complete medium and incubated in the 37 °C incubator until the cell density had reached 0.5 x 10^6/ml, which was measured using a haemocytometer. Cells were transferred to a T-75 cell culture flask and further passaged every two days to a final volume of 80 ml while maintaining the cell density.

2.3 Protein extraction

Protein samples were prepared from Jurkat T-cells by the means of ultra-sonication. After cell counting with a haemocytometer, Jurkat T-cells (1x 10^7) were transferred from a T-75 flask to a 50 ml centrifuge tube where they were collected by centrifugation at 600 x *g* for 3 min. The cell pellet was washed three times with ice-cold PBS and then resuspended in 1 ml of ice-cold PBS with 1 mM PMSF. The cells were lysed on ice using a Q700 sonicator (QSonica). The ultra-sonication program was set as follows: 30 kHz, 60 cycles, ON time per cycle = 3 seconds (s), OFF time per cycle = 10 s. The cell lysate was then centrifuged at 15,000 x *g*, 4 °C for 10 min to remove cell debris, and the supernatant was divided into aliquots (200 µl each) and stored at -80 °C for further use.

2.4 Protein quantitation

Protein samples, including BSA, lysozyme C, and protein extracts from Jurkat T-cells, were quantified by either the BCA or NanoDrop™ A_{280} assay. For BCA assays, a BSA protein solution was prepared at a concentration of 1000 µg/ml. Protein standards were prepared from the BSA solution through serial dilution with ddH_2O to a concentration of 0, 0.5, 5, 10, 20, or 30 µg/ml. Jurkat or individual protein samples were also diluted 100-, 250-, or 500-fold. The BCA assay was carried out on triplicates of the protein standards and the diluted samples (150 µl) in a 96-well plate according to the manufacturer's protocol. The absorbances of the protein standards and the diluted samples were then measured at 562 nm using a Varioskan LUX microplate reader (Thermo Fisher Scientific). A standard curve of protein concentration

was produced from the readings of the protein standards, based on which the concentration of each protein sample was determined.

For NanoDrop™ A_{280} assays, a buffer of choice (1 µl) was directly spotted onto a NanoDrop™ 2000 instrument (Thermo Fisher Scientific), and its absorbance at 280 nm was set as blank. Jurkat or individual protein samples (1 µl) were then added and the absorbance was measured to estimate protein concentration.

2.5 Negative selection of protein N-termini by NHS-Sepharose method

Individual protein samples were prepared in 20 mM Na_2CO_3 buffer (pH 8.5) at a concentration of 1 µg/µl. The amino (N)-terminal peptide of each protein was negatively selected using the "NHS-Sepharose method" as previously described (McDonald and Beynon, 2006). Briefly, 100 µg of lysozyme C or BSA was acetylated with 1 mg of sulfo-NHS acetate for 2 hours (h). The acetylated proteins were then incubated with 5 mg of Tris(2-aminoethyl)amine, polymer-bound for 1 h to quench the reaction. Then proteins were then purified from residual reactants by TCA (trichloroacetic acid) precipitation. In detail, 1/4 of the sample volume of 100 % (w/v) TCA solution was added, well mixed by vortexing, and incubated on ice for 1 h. The proteins were then precipitated by centrifugation at 15,000 x *g* for 15 min and the pellet was extensively washed three times with ice-cold diethyl ether. Finally, the pellet was resuspended in 20 mM Na_2HPO_4 (pH 7.5) to a protein concentration of 1 µg/µl, as measured by a NanoDrop™ A_{280} assay.

The purified proteins were denatured and reduced with 5 mM DTT for 1 h at 65 °C, alkylated with 15 mM iodoacetamide for 30 min in the dark, and finally digested with trypsin (for BSA) or Glu-C (for lysozyme C) at a 1:50 ratio (w/w, enzyme to protein) at 37 °C overnight. The digested samples (50 µg) were diluted with an equal volume of PBS (pH 7.5) and incubated with 100 µl of NHS-activated Sepharose twice, first at room temperature (RT) for 4 h then at 4 °C overnight, to negatively select N-terminal peptides.

2.6 Systematic refinement of NHS-Sepharose approach

Various experiment parameters were systematically and independently modified to optimise the NHS-Sepharose method: I. 100 µg of proteins were acetylated with 100 µg of sulfo-NHS acetate for 2 h, and then incubated with 10 mg of Tris(2-aminoethyl)amine, polymer-bound for 1 h; II. 50 µg of the digested samples (in PBS, pH 7.5) were incubated with 300 µl of NHS-activated magnetic beads at RT for 2 h; III. the digested samples were incubated with 100,

200, or 300 µl of NHS-activated Sepharose at RT for 4 h; IV. the digested samples were incubated with 100 µl of NHS-activated Sepharose at RT for 1, 2, 3, 4 or 8 h, or at 4 °C for 3, 6, 9, 12 or 24 h; V. the digested samples were incubated with 100 µl of NHS-activated Sepharose in PBS (pH 7.5, 8, or 8.5) at RT for 4 h; VI. the digested samples were incubated first with 8 M urea or 6 M guanidine HCl for 1 h then with 100 µl of NHS-activated Sepharose at RT for 4 h; VII. the digested samples were incubated with 200 µl of NHS-activated Sepharose twice (first at RT for 4 h then at 4 °C for 12 h), three times (two 4-h incubation at RT then at 4 °C for 12 h), or four times (two 4-h incubation at RT and two 12-h incubation at 4 °C).

Finally, BSA or lysozyme C was subjected to an optimised NHS-Sepharose protocol that integrated the above changes: 100 µg of the proteins were acetylated with 100 µg of sulfo-NHS acetate for 2 h and the reaction was quenched with 10 mg of Tris(2-aminoethyl)amine, polymer-bound for 1 h; following TCA precipitation, the acetylated proteins were tryptic digested as described above; 50 µg of the digested samples were incubated with 8 M of urea or 6 M guanidine HCl for 1 h, and then treated with 200 µl of NHS-activated Sepharose at pH 8.5 four times (two 4-h incubations at RT and two 12-h incubations at 4 °C).

2.7 Blocking of primary amines with SATA or CA

Protein extracts from Jurkat T-cells were first treated with DNase I at a ratio of 10 units/ml (enzyme to protein sample) at 37 °C for 1 h to remove genomic DNA. The reaction was quenched by incubating the samples with 5 mM EDTA (final concentration) at 65 °C for 10 min. The Jurkat protein samples were then buffer exchanged into PBS (pH 7.5) or 20 mM Na_2CO_3 buffer (pH 8.5) without protease inhibitors through ultrafiltration. In detail, the protein samples were added to a 3K (MWCO) centrifugal unit and centrifuged at 13,000 x *g* (4 °C) for 3 × 15 min, each with the addition of 500 µl PBS or Na_2CO_3 buffer, to complete buffer exchange. The protein samples (1 µg/µl) were incubated with SATA in PBS at a ratio of 50 nmol/µg (SATA to protein) for 1 h. The urea solution (8 M) was then added drop-wise to the SATA-modified proteins to dissolve any observed precipitates. Alternatively, the protein samples (1 µg/µl) were incubated with CA in in Na_2CO_3 buffer at a ratio of 20 nmol/µg (CA to protein) for 1 h. The modified proteins were further buffer exchanged into PBS (pH 7.5) by a second ultrafiltration step. Finally, the protein samples were digested with trypsin as described above.

2.8 Transamination

The N-terminal amino group of a protein or peptide was selectively converted into a carbonyl group via a transamination reaction. First, model protein samples (lysozyme C or BSA) were prepared by dissolving the protein powder in 0.1 M MES buffer (pH 6.0) with 4 M urea at a concentration of 2 µg/µl. The protein samples were then denatured, reduced, and alkylated as described previously prior to protein transamination. In contrast, Jurkat protein samples were denatured, reduced, and alkylated immediately after protein extraction. After solubilisation with 6 M guanidine HCl, the Jurkat proteins were buffer exchanged into 0.1 M MES buffer (pH 6.0) with 4 M guanidine HCl using Zeba™ desalting columns. In detail, the protein samples were added to the columns that had been conditioned with the said buffer. The columns were then centrifuged at 1,000 x *g* at 4 °C for 2 min to complete buffer exchange. The Jurkat proteins were diluted afterwards to a concentration of 2 µg/µl for protein transamination.

For peptide-level transamination, the Jurkat proteins were first digested with trypsin to yield Jurkat peptide samples at a concentration of 2 µg/µl (see section 2.12). In contrast, synthetic peptide samples (human ACTH or rat renin substrate) were prepared in a similar way to the model protein samples but without the addition of guanidine HCl. Both peptide samples were subjected to transamination as described below.

The protein or peptide samples (2 µg/µl) were incubated with an equal volume of the 2X "salt-free" reaction mixture (20 % (v/v) pyridine, 0.4 M glyoxylic acid, 12 mM $CuSO_4$) at RT for 2 h to complete transamination (Sonomura *et al.*, 2009a). This reaction condition was adopted in lieu of the 2X "salt-based" reaction mixture (6.6 M sodium acetate, 1 M sodium glyoxylate, 0.1 M $CuSO_4$, and 1 M acetic acid; Papanikos *et al.*, 2001), which produced a similar result for the synthetic peptides but was not compatible with the protein samples due to precipitation issues.

For further modification of the introduced carbonyl groups, the transaminated proteins were buffer exchanged into 20 mM phosphate buffer (pH 6.5) with 2 M guanidine HCl using Zeba™ columns. On the other hand, the transaminated peptides were purified using ZipTip® C18 pipette tips, vacuum dried using a SpeedVac™ concentrator (Thermo Fisher Scientific), and finally resuspended in the phosphate buffer with 2 M guanidine HCl.

2.9 Chemical tagging of the transaminated proteins or peptides

The newly formed N-terminal carbonyl group was selectively modified with $BnONH_2$, carbonyl-reactive biotins, or DNPH. For modification with $BnONH_2$, the transaminated proteins (1 µg/µl) were reacted with 50 mM $BnONH_2$ (final concentration) in 20 mM phosphate buffer with 2 M guanidine HCl (pH 6.5) at 37 °C for 2 h. The modified proteins were buffer exchanged into PBS (pH 7.5) using Zeba™ columns. For biotinylation, the transaminated proteins or peptides (1 µg/µl) were reacted with 5 mM hydrazide-biotin, alkoxyamine-PEG_4-biotin, or alkoxyamine-PEG_4-SS-PEG_4-biotin (final concentration) under the same conditions, and then buffer exchanged using Zeba™ columns. For modification with DNPH, the protein samples were precipitated by the TCA method immediately after transamination. DNPH (0.2 %, w/v) in 2 N HCl was added to the pellet and tubes were then agitated for 15 min in the dark. The modified proteins were precipitated again with TCA, and then dissolved in 6 M guanidine HCl to a concentration of 1 µg/µl (measured by NanoDrop™ A_{280} assays). In contrast, the transaminated peptides were modified with $BnONH_2$ or carbonyl-reactive biotins under the same conditions but then directly desalted using ZipTip® C18 pipette tips for liquid chromatography–tandem mass spectrometry (LC-MS/MS).

2.10 Enrichment of the biotinylated proteins

The biotinylated proteins (1 µg/µl in PBS, pH 7.5) were subjected to affinity purification (AP) together with three control groups: native-state proteins (i.e. without either transamination or biotinylation), transaminated proteins without biotinylation, and proteins only treated with biotin. 50 µl of NeutrAvidin or monomeric avidin agarose beads (50 % slurry in storage solution) were first conditioned according to the manufacturer's instructions. The beads were then ready to receive one of the four samples, diluted with PBST (0.1 % (v/v) Tween-20 in PBS, pH 7.5) to 500 µl, and incubated at RT with end-over-end turning for 1 h. The beads were collected by centrifugation at 500 x *g* for 2 min, washed first with 3 × 200 µl PBST, and then washed with 3 × 200 µl PBS. The NeutrAvidin beads were boiled in 50 µl of 1X SDS sample buffer for 5 min to elute proteins bound to the beads. In contrast, the monomeric avidin beads were treated with 40 µl of 0.1 M glycine (pH 2.8) to elute proteins, which was immediately neutralised with 10 µl of 1 M Tris HCl (pH 8.5). An aliquot of each sample was added with 5X SDS sample buffer and boiled for 10 min prior to sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE).

2.11 Gel electrophoresis and immunoblotting

To check genomic DNA removal, an aliquot of Jurkat protein samples were mixed with 1/5 volume of DNA gel loading dye (6X) and loaded onto a 1 % (w/v) agarose gel. Gel electrophoresis was performed in TAE buffer at 100 V for 1 h. The agarose gels were stained with 1X GelRed®. Finally, DNA bands were visualised under ultraviolet (UV) light using a Gel Doc™ XR+ imager (Bio-Rad).

For SDS-PAGE, protein samples were prepared in 1X SDS sample buffer and boiled for 10 min. 15 µl of each sample (0.5 µg/µl) was loaded onto a 10 % precast gel, together with protein standards. Gel electrophoresis was performed in 3-(*N*-Morpholino)propanesulfonic acid (MOPS) running buffer (2.5 mM MOPS, 2.5 mM Tris, 0.005 % SDS, 0.05 mM EDTA, pH 7.7, supplemented with 1/400 volume of NuPAGE™ antioxidant) at 200 V for 50 min. The gels were stained with Colloidal Blue for 3 h and then destained with ddH_2O overnight.

Western blotting was employed to inspect the chemical tagging step before AP. After SDS-PAGE, proteins were transferred from the gel to a nitrocellulose membrane in transfer buffer (25 mM Tris, 192 mM glycine, 20 % (v/v) methanol, pH 8.3) on ice at 100 V for 1 h. Proteins transferred to the membrane were stained with Ponceau S to assess transfer efficiency, and then destained with PBST. The membrane was incubated in blocking solution (5 % (w/v) non-fat milk in PBST) at RT for 1 h, and then blotted with biotin- or dinitrophenol-specific primary antibodies (rabbit, 1:1,000 diluted in blocking solution) at 4 °C overnight. Following 3 x 15-min washing with PBST, the membrane was incubated with IRDye® 680 LT anti-rabbit IgG secondary antibody (donkey) at RT for 1 h, and then washed three times with PBST for 15 min each. Finally the membrane was scanned by an Odyssey® infrared imaging system (LI-COR) to detect proteins tagged with biotin.

Additionally, dot blotting was employed to rapidly identify proteins tagged with biotin or dinitrophenol groups. 5 µl of the tagged protein sample (1 µg/µl) was directly spotted onto a nitrocellulose membrane, together with three control samples as described previously. Sample loading was checked using Ponceau S stain before the membrane was subjected to the same immunoblotting procedure mentioned above.

2.12 Sample preparation

Protein samples were subjected to the preparation procedure as described previously (if not already done): proteins were denatured and reduced with 5 mM DTT at 65 °C for 1 h, and then alkylated with 15 mM iodoacetamide for 30 min in the dark. The samples were then diluted with a buffer of choice (e.g. PBS), to lower the concentration of guanidine HCl to < 1 M, and digested with trypsin or Glu-C at a 1:50 ratio (w/w, enzyme to protein) at 37 °C overnight. Protein digests were lyophilised afterwards in a SpeedVac™ concentrator to complete dryness and resuspended in 0.5 % (v/v) formic acid for ZipTip® desalting.

2.13 LC-MS/MS

Protein digests and peptide samples (100 ng) were analysed using an EASY-nLC 1000 system coupled online to an LTQ-Orbitrap Elite™ (Thermo Fisher Scientific). Synthetic peptides and the digests of individual proteins were first loaded onto an Acclaim™ PepMap™ C18 trap column (0.075 mm × 150 mm, particle size = 3 µm, pore size = 100 Å, Thermo Fisher Scientific), and then separated on an analytical column (Acclaim™ PepMap™ C18, particle size = 2 µm) using a 60-min binary gradient consisting of 0.1 % (v/v) formic acid in ddH_2O (solvent A) and 0.1 % (v/v) formic acid in acetonitrile (solvent B). The gradient was as follows: 0 – 5 % solvent B in 1 min, 5 – 25 % B over 45 min, 25 – 40 % B in 5 min, 40 – 95 % B in 2.5 min, and finally held at 95 % B for 6.5 min. In contrast, Jurkat protein digests were eluted from the trap column and analysed on a MonoCap™ C18 HighResolution 3000 column (0.1 mm x 3,000 mm, through-pore size = 2 µm, mesopore size = 15 nm, GL Sciences) at a flow rate of 300 nl/min with the following gradient: 0 – 10 % solvent B in 5 min, 10 – 45 % B in 360 min, and 45 – 90 % B in 10 min.

Data acquisition was set at data-dependent mode: in each cycle, one MS1 scan (full mass spectrum) was recorded with 10 MS2 scans (tandem mass spectra) corresponding to the 10 most intense precursor ions. Full mass spectra were acquired at a resolution (*R*) of 60,000 from 150 to 2,000 *m/z*. Precursor ions were fragmented by Higher-energy Collisional Dissociation (HCD) to produce tandem mass spectra. Dynamic exclusion was set to 30 s, and +1 charged ions were rejected for fragmentation. The normalised collision energy for HCD was set to 35 %, the isolation window was 2.0 *m/z*, and the activation time was 0.1 ms. Tandem mass spectra were acquired at *R* = 15,000 (*m/z* range = 150 – 2000) with preview mode enabled.

2.14 Data analysis

Orbitrap RAW data were manually inspected using Xcalibur Qual Browser (Thermo Fisher Scientific) before being converted into peak lists by Mascot Distiller (Matrix Science) in Mascot generic format (MGF). These data were then searched against the decoy sequences of synthetic peptides/individual proteins (manually entered), or those from the Swiss-Prot database (taxonomy = *Homo sapiens*), using Mascot Daemon 2.4. Protein search settings were as follows: trypsin or Glu-C protease specificity with up to two missed cleavages for protein digests; "None Cutting Enzyme" for peptides; cysteine (C) carbamidomethylation (+57.02 Da) as fixed modification; methionine (M) oxidation (+15.99 Da), N-terminal/lysine (Nt/K) acetylation (+42.01 Da), Nt/K modification with SATA (+115.99 Da) or CA (+ 112.02 Da), Nt-transamination (-1.03 Da), N-terminal tagging with $BnONH_2$ (+104.03 Da), hydrazide-biotin (+239.10 Da), alkoxyamine-PEG_4-biotin (+415.18 Da), or alkoxyamine-PEG_4-SS-PEG_4-biotin (+825.33 Da) as variable modifications (depending on the subject of each study); charge states were set as +2 and +3 only; peptide and MS/MS tolerance were set at 10 ppm (parts per million) and 0.5 Da, respectively.

In searching for unsuspected modifications of synthetic peptides/model proteins due to transamination (see sections 4.2.2 & 4.2.3), automatic error-tolerant searches were also performed in Mascot Daemon 2.4 as specified by the software provider. On the other hand, protease specificity was set to semi-trypsin in an effort to identify *neo*-N-termini in Jurkat protein samples (see section 5.3).

All the RAW, MGF, and Mascot search results were archived in the local network storage system, ReadyNAS NV+ v2 (NETGEAR). The data of Jurkat shotgun proteomics and Jurkat peptide transamination have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Vizcaíno *et al.*, 2015) with the dataset identifier PXD009340 and PXD009427, respectively.

2.15 Data mining

Mascot search results were only accepted if the false discovery rate (FDR) was < 5 %. In general, Mascot data were further processed in R to select significant peptide hits with a peptide expectation (*E*)-value ≤ 0.05, and to remove duplicate peptide hits. "One-hit wonders" (proteins identified by a single peptide) were further removed from the protein identification data in order to comply with the "two-peptide rule" (i.e. a protein msut be identified on the basis of ≥ 2 significant peptide hits; Bradshaw *et al.*, 2006). For the

transaminated Jurkat peptides, peptide hits with an *E*-value > 0.05 and duplicate hits were both removed but one-hit wonders were included in the initial results. This decision was based on the claim by Gupta and Pevzner (2009) that excluding one-hit wonders actually lowers the sensitivity of proteomics and that a large proportion of one-hit wonders are indeed expressed. Due to the high mass accuracy and resolution of LTQ-Orbitrap Elite™, peptide hits with an *E*-value ≤ 0.05 are highly reliable. In error-tolerant searches, a peptide hit was only accepted if the peptide score was no less than 13, which corresponds to a *P*-value ≤ 0.05.

The final data of peptide detection and protein identification were statistically analysed and visualised using R. In particular, the isoelectric point (pI) and hydrophobicity index of Jurkat peptides were computed in R using the *Peptides* package (Osorio *et al.*, 2015). Gene Ontology (GO) term enrichment and pathway analysis were performed using the DAVID Bioinformatics Resources v6.8 (Huang *et al.*, 2008).

Chapter 3. Critical assessment of the NHS-Sepharose approach for recovery of N-terminal peptides

3.1 Introduction

The identification of either protein amino (N)-termini or proteolytic processing events increasingly relies on the techniques of positional proteomics, which target peptides with defined positions after protease digestion in standard shotgun proteomics (Eckhard *et al.*, 2016). The two major categories of peptides investigated by such techniques are protein N- and carboxyl (C)-termini. At present, the focus is primarily on protein N-termini since primary amines are more reactive than carboxyl groups (Plasman *et al.*, 2013). A group of techniques have been developed to selectively recover such peptides through the semi-unique chemistry of primary amines. Such techniques (also known as N-terminalomics) can be further divided into two types of strategy: "negative selection", including combined fractional diagonal chromatography (COFRADIC; Staes *et al.*, 2011) and N-terminal amine isotopic labeling of substrates (N-TAILS; Kleifeld *et al.*, 2010), and "positive selection", such as the Subtiligase method (Mahrus *et al.*, 2008).

At the University of Liverpool, the Beynon group has also developed a negative selection strategy that removes internal peptides after proteolysis, thereby allowing the capture of protein N-termini (McDonald *et al.*, 2005). This promising method employs an acetylation reaction that blocks both α-amino groups (protein N-termini) and the ε-amino groups of lysine (K) residues on intact proteins. Following protease digestion of the acetylated proteins, the method uses *N*-hydroxysuccinimide (NHS)-activated Sepharose (an amine-specific coupling reagent in immobilised form) to remove peptides with free N-termini from the mixture (Figure 3.1). However, this method (referred to as the "NHS-Sepharose" approach hereafter) has not been as fully exploited in contrast to other negative selection techniques. As suggested by Dr Xumin Zhang, complete removal of internal peptides could not be achieved using the NHS-Sepharose approach (2016, personal communication). In light of this statement, we aimed to assess the feasibility of this method using commercially available proteins (see section 3.2.1).

It was further envisaged that the experimental conditions used in the NHS-Sepharose approach could be refined to improve the efficiency of peptide removal, which would in turn promote a wider use of this method. Seven experiment parameters to be refined were empirically determined. First, it was suspected that the acetylation reagent, when in a large

molar excess, might carry over to the subsequent reaction steps (e.g. protease digestion) and thus negatively impact the outcome of peptide removal. Next, the coupling reagent of choice is integral to this method and often requires fine-tuning for specific applications (e.g. Mejía-Manzano *et al.*, 2016). Furthermore, the pH, temperature, incubation times, reagent quantity, and clean-up method are also conceivable targets for refinement. For instance, the specificity of the coupling reaction toward different amino groups can be altered by lowering the pH from 8.5 to 6.5 (Selo *et al.*, 1996).

Therefore, the present study also aimed to assess the scope for improvement by systematically refining seven experiment parameters of the NHS-Sepharose approach: I. the amount of acetylation reagents; II. the type of coupling reagents; III. the amount of coupling reagents; IV. the duration of the coupling reaction; V. coupling reaction pH; VI. the repeat number of the coupling reaction; VII. the use of chaotropic agents (to account for any hydrophobic effect, see section 3.2.2). An optimised protocol of this method was further coupled to alternative reactions with primary amines. Finally, a shotgun proteomic study was conducted to assess the feasibility of the optimised method (with alterative reactions) on a proteome-wide scale (see section 3.2.3).

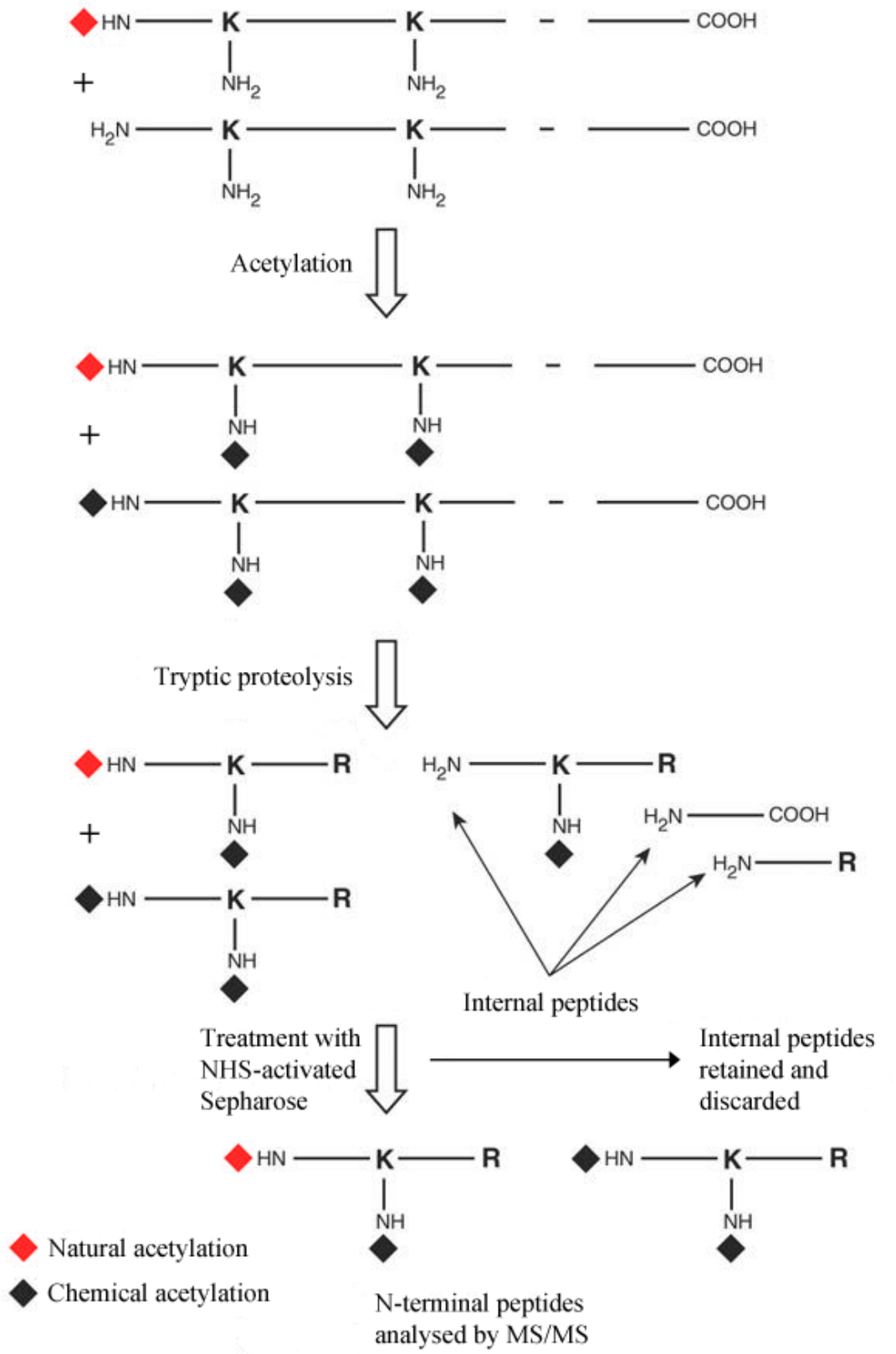


Figure 3.1 Schematic of the NHS-Sepharose approach, which employs acetylation of primary amines and amine-reactive resin (i.e. NHS-activated Sepharose) to negatively select the N-terminal peptides of proteins after proteolysis (modified from McDonald *et al.*, 2005). Furthermore, with the use of isotopic labels this approach allows discrimination between naturally and chemically acetylated N-termini. NHS: *N*-hydroxysuccinimide; MS/MS: tandem mass spectrometry.

3.2.1 Implementation of the NHS-Sepharose approach

The experimental protocol for the NHS-Sepharose approach was initially tested on chicken egg-white lysozyme (lysozyme C, Swiss-Prot ID: LYSC_CHICK/P00698) in conjunction with Glu-C protease digestion. Lysozyme C is a bacteriolytic enzyme consisted of 129 amino acid residues (Figure 3.2). The protein precursor contains an 18-residue signal peptide, which is cleaved off after protein synthesis; as a result the mature protein starts with an N-terminal K residue (Canfield, 1963).

Figure 3.2 Protein sequence of the mature lysozyme C. The sequence shown in red was matched to the experimental data by Mascot database search (see Table 3.1).

A test experiment was first performed to identify the predicted N-terminal peptide after Glu-C digestion. The peptides were subjected to liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis. Mascot Distiller software was then utilised to convert the collected data to peak lists, which were searched against a decoy version of the protein sequence of lysozyme C using Mascot Daemon. As shown in Table 3.1, there were 15 Glu-C generated peptides of lysozyme C identified with high confidence (*E*-value ≤ 0.05), leading to 100 % protein sequence coverage (Figure 3.2).

The authentic N-terminal peptide of lysozyme C was identified as lysine – valine – glycine – arginine – cysteine (carbamidomethylated) – glutamate (KVFGRCE) after Glu-C digestion. The N-terminal peptide eluted at 12.75 minute (min) during LC separation (Figure 3.3). It was detected as a precursor ion at m/z = 448.23 with two positive charges (z = 2), which corresponds to a molecular weight (MW) of 894.44 Da (Figure 3.4). The sequence of this peptide was obtained by matching the tandem mass spectrum for the precursor ion to the *in silico* predicted m/z values for all possible fragment ions. The tandem mass spectrum was obtained via Higher-energy Collisional Dissociation (HCD), which produces predominantly the *b*- and *y*-series fragment ions. HCD-produced fragment ions can be assigned to their corresponding amino acid residues by fitting their m/z to the predicted value for each *b*- or *y*-ion.

Table 3.1 Identification of Glu-C generated peptides of lysozyme C by a Mascot database search[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value
KVFGRCE (1 – 7) + Carbamidomethyl (C)	448.2275	894.4382	36	0.00023
KVFGRCELAAAMKRHGLD (1 – 18) + Carbamidomethyl (C)	687.0315	2058.0615	33	0.0005
LAAAMKRHGLD (8 – 18)	591.8257	1181.6339	45	2.9E-5
LAAAMKRHGLDNYRGYSLGNWVCAAAKFE (8 – 35) + Carbamidomethyl (C)	1066.8644	3197.5651	33	0.00045
NYRGYSLGNWVCAAAKFE (19 – 35) + Carbamidomethyl (C)	1017.9816	2033.9418	61	8.7E-7
SNFNTQATNRNTD (36 – 48)	741.8341	1481.6495	83	5.6E-9
SNFNTQATNRNTDGSTD (36 – 52)	921.8981	1841.7776	51	7.6E-6
GSTDYGILQINSRWWCND (49 – 66) + Carbamidomethyl (C)	1092.9968	2183.9695	17	0.021
YGILQINSRWWCND (53 – 66) + Carbamidomethyl (C)	912.9313	1823.8413	59	1.4E-6
GRTPGSRNLCNIPCSALLSSD (67 – 87) + 2 Carbamidomethyl (C)	759.0386	2274.0845	48	1.4E-5
GRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSD (67 – 101) + 3 Carbamidomethyl (C)	941.2290	3760.8658	59	1.4E-6
ITASVNCAKKIVSD (88 – 101) + Carbamidomethyl (C)	753.4059	1504.7919	46	2.5E-5
ITASVNCAKKIVSDGNGMNAWVAWRNRCKGTD (88 – 119) + 2 Carbamidomethyl (C)	895.6915	3578.7293	88	1.6E-9
GNGMNAWVAWRNRCKGTD (102 – 119) + Carbamidomethyl (C)	698.3235	2091.9479	50	9.4E-6
VQAWIRGCRL (120 – 129) + Carbamidomethyl (C)	629.8483	1257.6764	21	0.0078

[a] Orbitrap RAW data were processed and searched in Mascot twice, with cysteine (C) carbamidomethylation set as either a fixed or variable modification. Both Mascot searches produced the same result, which was then edited to show a single PSM representing each peptide identified with high confidence (*E*-value ≤ 0.05). The doubly charged ion at *m/z* = 448.23 (see Fig. 3.4) was matched to the carbamidomethylated N-terminal peptide KVFGRCE. PSM: peptide-spectrum match; *m*/z: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value.

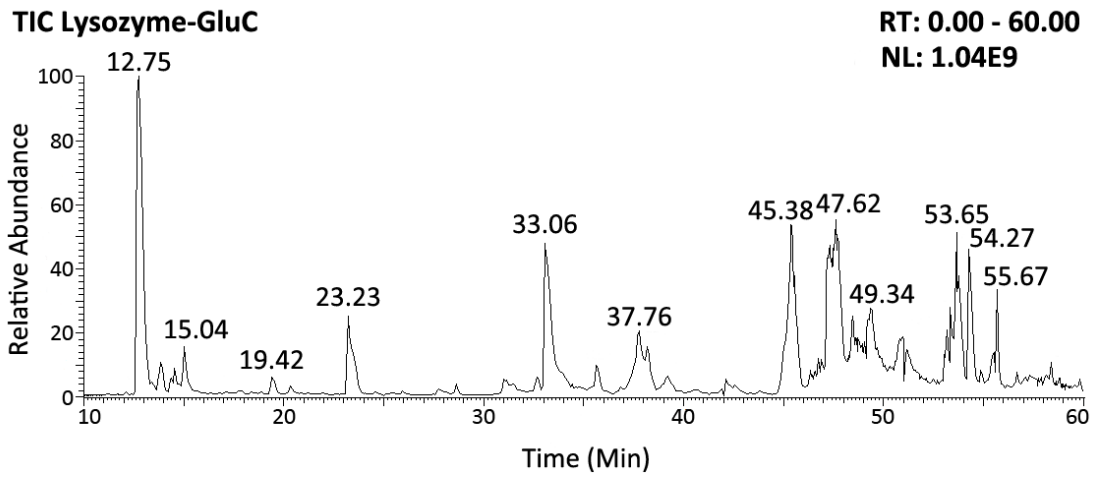


Figure 3.3 Total ion current chromatogram (TIC) of lysozyme C peptides produced by Glu-C protease digestion and separated on a C18 column (Acclaim™ PepMap™) for mass spectrometric analysis. Peaks in the chromatogram correspond to putative proteolytic peptides. The peak eluting at *RT* = 12.75 min is analysed below (Fig. 3.4). *RT*: retention time; NL: normalised intensity level; min: minute.

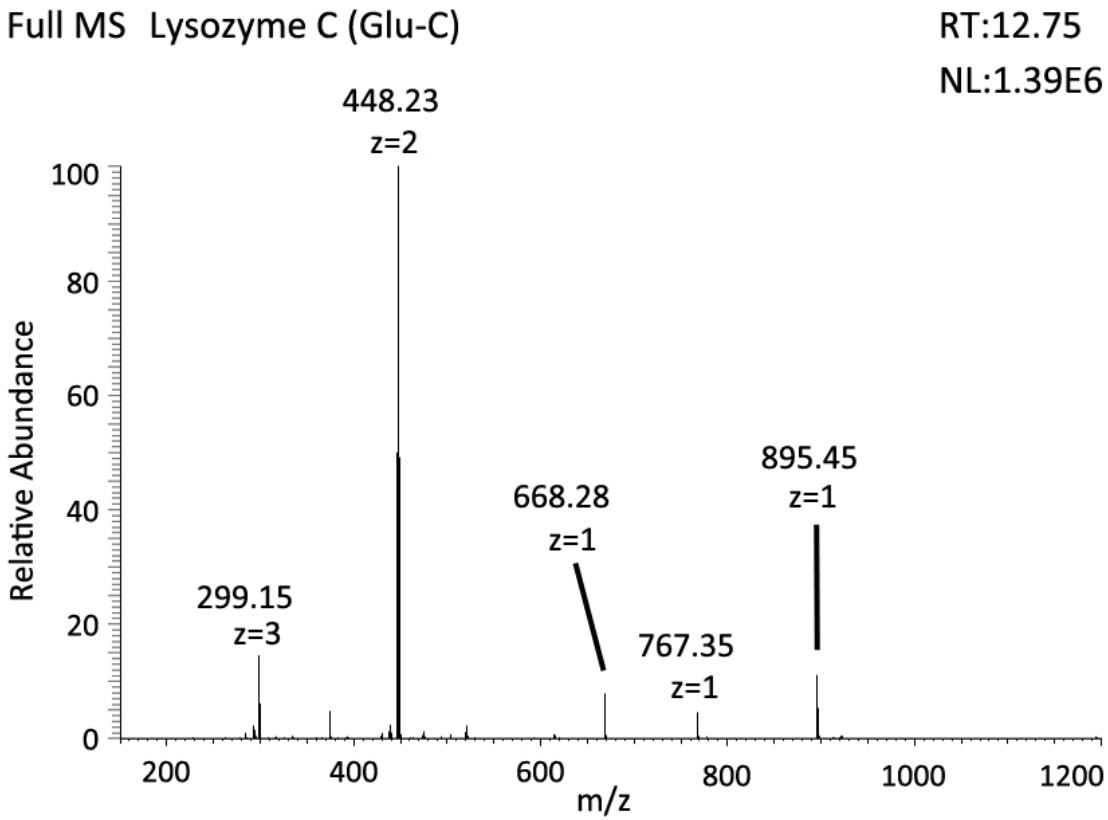


Figure 3.4 Full mass spectrum of the peptide eluting at *RT* = 12.75 min. The major peak in the graph is a doubly charged ion (*z* = 2) with a mass-to-charge value (*m*/*z*) = 448.23. *RT*: retention time; NL: normalised intensity level; min: minute.

As shown in Figure 3.5 and Table 3.2, this tandem mass spectrum contains five matched *y*-ions ($y_2 - y_6$), which correspond to the carboxyl parts of the peptide KVFGRCE. The matched b_1-ion was excluded as the b_1-ion of an unmodified N-terminal residue is never seen (Maleknia and Johnson, 2011). A unique feature of this peptide is the presence of both an α-amino group and an ε-amino group on the N-terminal K residue, without endogenous N-terminal (Nt)-acetylation. In principle, this chemical property requires the addition of two acetyl groups ($-COCH_3$) in order to fully block the protein N-terminus.

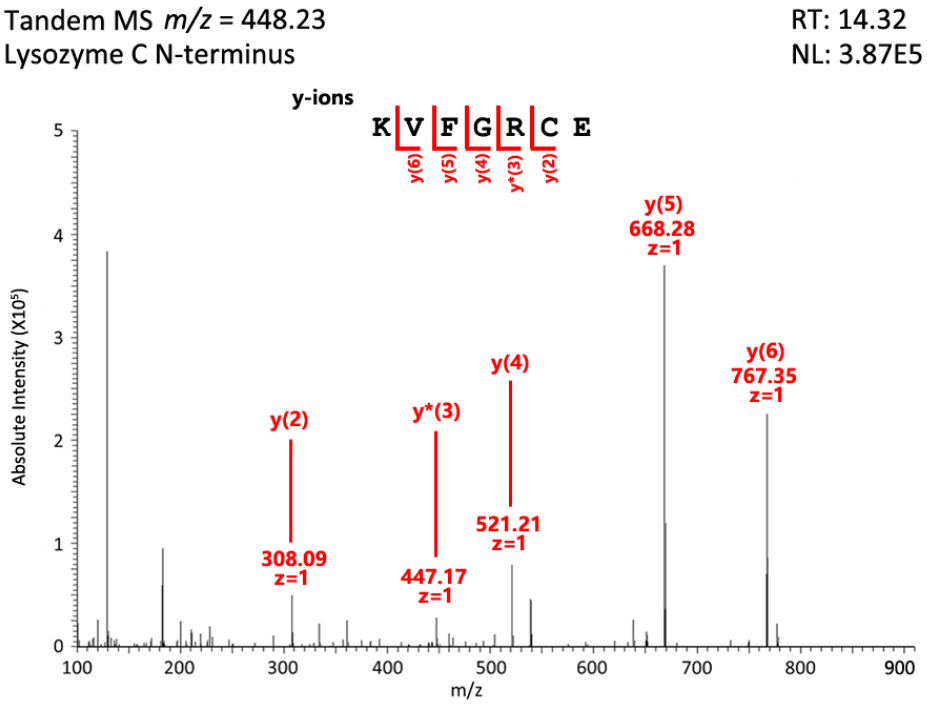


Figure 3.5 Tandem mass spectrum of the precursor ion at *m/z* = 448.23 (*z* = 2). This ion corresponds to the N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE). The spectrum was matched to this peptide based on the *in silico* prediction of fragment ions (shown in Table 3.2). *RT*: retention time; NL: normalised intensity level.

Table 3.2 Fragment ions (*in silico* predicted) of the peptide KVFGRCE. Fragment ions matched to the experimental data are shown in red[a].

#	b	b^{++}	b*	b^{*++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	129.1022	65.0548	112.0757	56.5415	K							7
2	228.1707	114.5890	211.1441	106.0757	V	767.3505	384.1789	750.3239	375.6656	749.3399	375.1736	6
3	375.2391	188.1232	358.2125	179.6099	F	668.2821	334.6447	651.2555	326.1314	650.2715	325.6394	5
4	432.2605	216.6339	415.2340	208.1206	G	521.2137	261.1105	504.1871	252.5972	503.2031	252.1052	4
5	588.3616	294.6845	571.3351	286.1712	R	464.1922	232.5997	447.1656	224.0865	446.1816	223.5945	3
6	748.3923	374.6998	731.3657	366.1865	C	308.0911	154.5492			290.0805	145.5439	2
7					E	148.0604	74.5339			130.0499	65.5286	1

[a] Matched b_1 ion (*m/z* = 129.1022) was excluded since the unmodified N-terminal b_1 ion is never detected and the matched ion may correspond to an iminium ion of the lysine (K) residue.

Having experimentally established the N-terminal peptide of lysozyme C, the next step was to examine the effect of primary amine acetylation. Accordingly, the protein was incubated with the acetylation reagent (sulfo-NHS acetate) for 2 hours (h) prior to quenching with Tris(2-aminoethyl)amine, polymer bound. Following precipitation, the protein was denatured, reduced, alkylated, and then proteolysed with Glu-C. Finally, the proteolytic peptides were analysed by LC-MS/MS. A control experiment was performed where lysozyme C was digested with Glu-C without acetylation.

As shown in Table 3.3 (and later in Figure 3.9), 86 % of the lysozyme C sequence was identified in the acetylation experiment based on the high-confidence (*E*-value ≤ 0.05) match of 12 peptides generated by Glu-C digestion. The acetylated N-terminal peptide was identified as a doubly charged ion at m/z = 490.23, which corresponds to a MW of 978.46 Da (Figure 3.6). By comparing the results between the control and acetylation experiments, the N-terminal peptide KVFGRCE was shown to exhibit a mass shift of +84.02 Da (from 894.44 to 978.46 Da) after acetylation. As the MW of an acetyl group is 42.01 Da, this mass shift corresponds to the addition of two acetyl groups (Table 3.3).

Table 3.3 also shows that the vast majority of primary amines on lysozyme C were acetylated after the treatment with sulfo-NHS acetate. The N-terminal peptide was only identified in the form of single or double acetylation on the N-terminal K residue. This result suggests a high efficiency of acetylation at the N-terminus of lysozyme C since peptide ions with the free N-terminus were not detected. On the other hand, most of the internal peptides of lysozyme C were susceptible to acetylation if a K residue was present. The only exception was the internal peptide ITAVNCAKKIVSD that contains two adjacent K residues. Only one of the K residues in this peptide was acetylated, suggesting a negative effect of the adjacent K residue on acetylation.

Table 3.3 Glu-C generated peptides of acetylated lysozyme C as identified by a Mascot database search[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value
KVFGRCE (1 – 7) + Acetyl (N-term); Carbamidomethyl (C)	469.2306	936.4487	16	0.023
KVFGRCE (1 – 7) + Acetyl (N-term); Acetyl (K); Carbamidomethyl (C)	490.2346	978.4593	21	0.0083
KVFGRCELAAAMKRHGLD (1 – 18) + Acetyl (N-term); 2 Acetyl (K); Carbamidomethyl (C)	729.0373	2184.0932	50	1.0E-5
LAAAMKRHGLD (8 – 18) + Acetyl (K)	408.8887	1223.0444	31	0.00074
LAAAMKRHGLD (8 – 18) + Acetyl (K); Oxidation (M)	400.2178	1239.6394	25	0.0031
LAAAMKRHGLDNYRGYSLGNWVCAAAKFE (8 – 35) + 2 Acetyl (K); Carbamidomethyl (C)	1094.8626	3281.5862	14	0.035
NYRGYSLGNWVCAAAKFE (19 – 35) + Acetyl (K); Carbamidomethyl (C)	1038.9734	2075.9523	20	0.0092
SNFNTQATNRNTD (36 – 48)	741.8313	1481.6495	69	1.3E-7
SNFNTQATNRNTDGSTD (36 – 52)	921.8875	1841.7776	60	9.2E-7
YGILQINSRWWCND (53 – 66) + Carbamidomethyl (C)	912.9254	1823.8413	37	0.00022
GRTPGSRNLCNIPCSALLSSD (67 – 87) +2 Carbamidomethyl (C)	759.0316	2274.0845	53	4.7E-6
ITASVNCAKKIVSD (88 – 101) + Acetyl (K); Carbamidomethyl (C)	774.4073	1546.8025	14	0.037
GNGMNAWVAWRNRCKGTD (102 – 119) + Acetyl (K); Carbamidomethyl (C)	712.3232	2133.9585	22	0.0057
VQAWIRGCRL (120 – 129) + Carbamidomethyl (C)	629.8416	1257.6764	21	0.0078

[a] Orbitrap RAW data were processed and searched in Mascot twice, with cysteine (C) carbamidomethylation set as either a fixed or variable modification. Both Mascot searches produced the same result, which was then edited to show a single PSM representing each peptide identified with high confidence (*E*-value ≤ 0.05). PSM: peptide-spectrum match; *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value.

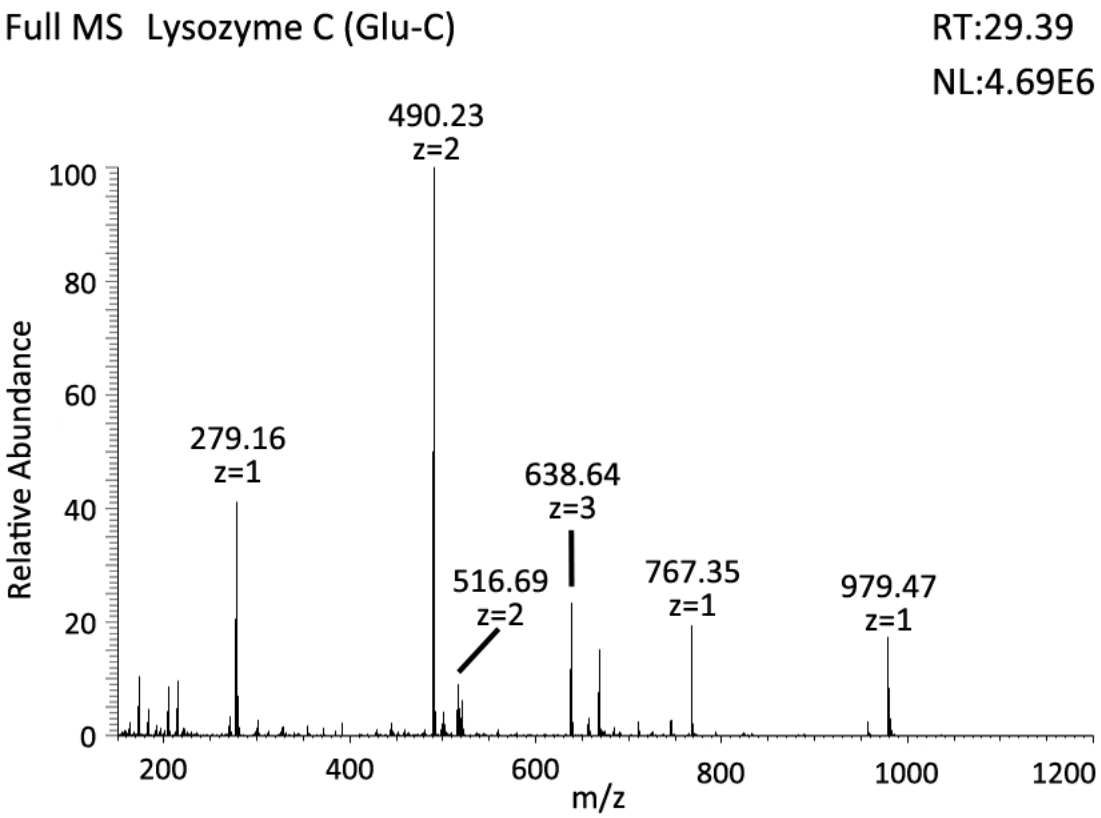


Figure 3.6 Full mass spectrum of the peak at *RT* = 29.39 min, which contained peptides from acetylated lysozyme C. The doubly charged ion at *m*/*z* = 490.23 is the major peak in this graph, which was derived from the acetylated N-terminal peptide of lysozyme C (see Fig. 3.7). *RT*: retention time; NL: normalised intensity level; *m*/*z*: mass-to-charge ratio; min: minute.

As shown in Figure 3.7, the tandem mass spectrum for the precursor ion at *m*/*z* = 490.23 contains five *y*-ions (y_2 – y_6) that were matched to the *in silico* predicted values (Table 3.4). The *m*/*z* values of these five *y*-ions were identical to those derived from the native N-terminal peptide previously shown in Figure 3.5 and Table 3.2. Hence it was inferred that only the N-terminal residue was modified by acetylation, as N-terminal modifications should not change the *m*/*z* of *y*-ions. In addition, only the N-terminal K residue in this peptide possesses free amino groups that are susceptible to acetylation. Two acetyl groups were added to the N-terminal residue instead of just one due to the fact that the N-terminal K residue contains two free amino groups. One of them corresponds to the α-amino group at the N-terminus while the other corresponds to the ε-amino group present on the side chain of a K residue, which happens to be the N-terminal amino acid residue of the mature lysozyme C. In conclusion, acetylation can be employed to block the N-terminus of lysozyme C and K side chains.

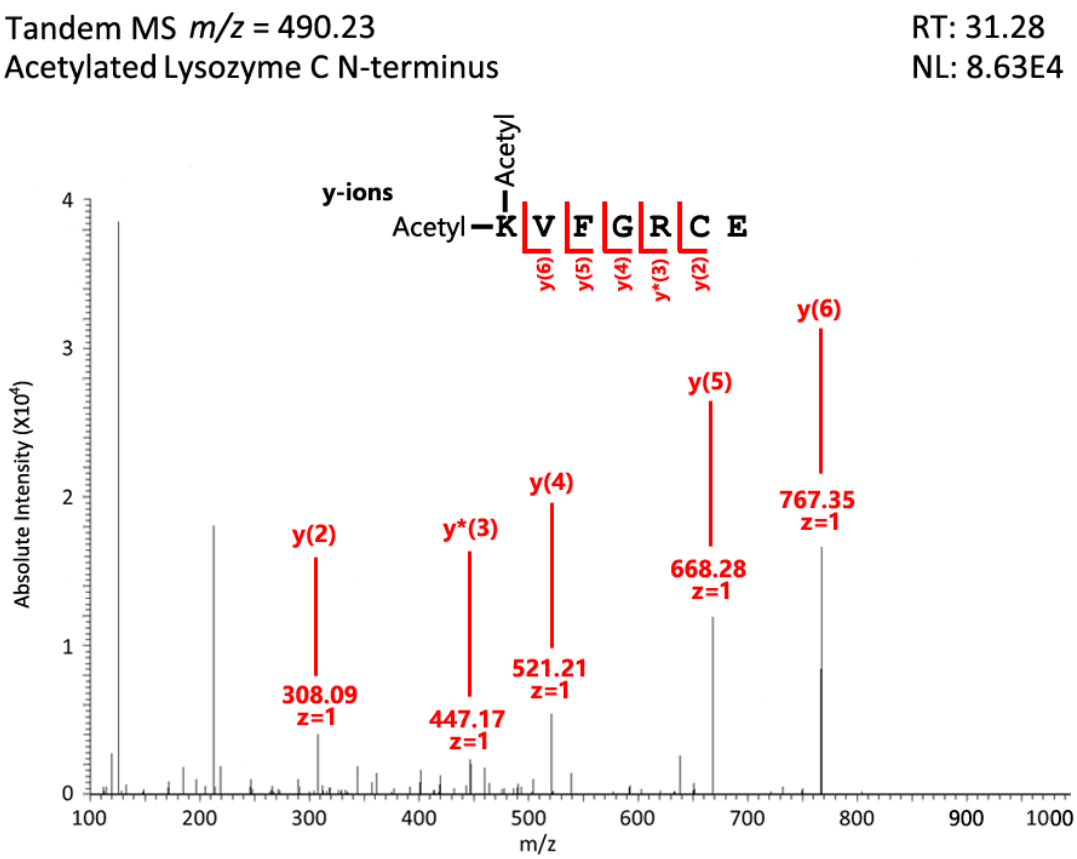


Figure 3.7 Tandem mass spectrum of the precursor ion at *m/z* = 490.23 (*z* = 2). This ion corresponds to the acetylated N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE). The spectrum was matched to this peptide based on the *in silico* prediction of fragment ions (shown in Table 3.4). *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio.

Table 3.4 Fragment ions (*in silico* predicted) of the acetylated peptide KVFGRCE. Fragment ions matched to the experimental data are shown in red.

#	b	b^{++}	b*	b^{*++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	213.1234	107.0653	196.0968	98.5520	K							7
2	312.1918	156.5995	295.1652	148.0863	V	***767.3505***	384.1789	750.3239	375.6656	749.3399	375.1736	6
3	459.2602	230.1337	442.2336	221.6205	F	***668.2821***	334.6447	651.2555	326.1314	650.2715	325.6394	5
4	516.2817	258.6445	499.2551	250.1312	G	***521.2137***	261.1105	504.1871	252.5972	503.2031	252.1052	4
5	672.3828	336.6950	655.3562	328.1817	R	464.1922	232.5997	447.1656	224.0865	446.1816	223.5945	3
6	832.4134	416.7103	815.3869	408.1971	C	***308.0911***	154.5492			290.0805	145.5439	2
7					E	148.0604	74.5339			130.0499	65.5286	1

Having established the efficiency of primary amine acetylation, the next step was to examine the efficiency of enrichment for the acetylated N-terminal peptide after removal of Glu-C generated internal peptides. Such peptides possess a free α-amino group at their N-termini, and in theory can be removed using amine-reactive chemistry. The reagent of choice was NHS-activated Sepharose, which is an immobilised amine-reactive resin. The hypothesis of this experiment was that the acetylated N-terminal peptide of lysozyme C would remain detectable by LC-MS/MS after treating the entire pool of Glu-C generated peptides with this resin; in contrast, other internal peptides would be removed by the resin and hence undetectable.

The N-terminal peptide enrichment was carried out according to McDonald and Beynon (2006). Following acetylation of lysozyme C and treatment of Glu-C generated peptides with NHS-activated Sepharose, the supernatant was collected, desalted, and analysed by LC-MS/MS. By comparing the base peak chromatograms between the control and treated samples, the vast majority of the peptides from acetylated lysozyme C have diminished after the treatment with NHS-activated Sepharose (Figure 3.8). As shown by Mascot search results, 86 % of the lysozyme C sequence was identified in the control sample, with 11 high-confidence peptide matches (Table 3.5). The number of peptide matches decreased to 3 in the treated sample, constituting only 18 % of protein sequence coverage (Figure 3.9).

The three identified peptides were the acetylated N-terminal peptide (KVFGRCE, with two acetyl groups) and two internal peptides (SNFNTQATNRNTD and SNFNTQATNRNTDGSTD) that both started with the same amino acid residue (S36). Repeated identification of such peptides suggests that it was either of high abundance due to highly efficient proteolysis by Glu-C, or ionised particularly efficiently in the electrospray process. Alternatively, the two internal peptides might have “masked” their N-termini or bind non-covalently to the authentic N-terminal peptide (KVFGRCE).

Due to the presence of internal peptides when the original NHS-Sepharose protocol was followed, the conclusion is as follows: the NHS-Sepharose approach is capable of removing the vast majority of internal peptides after protease digestion of acetylated proteins; however certain internal peptides may remain with the authentic N-termini. Without prior knowledge, these internal peptides may be misidentified as protein N-termini. Therefore, this approach requires further improvement or modification before it can be employed for systematic identification of protein N-termini or protease substrates.

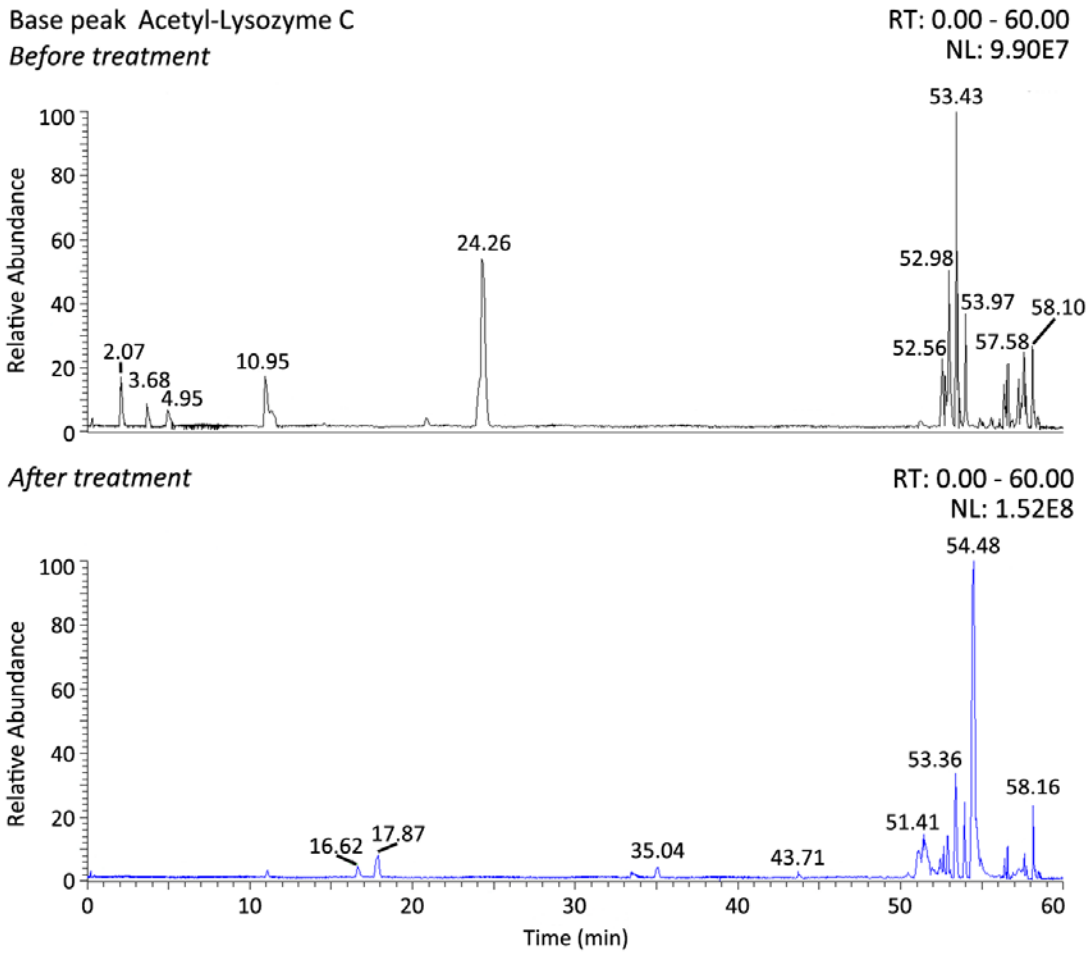


Figure 3.8 Comparison of the base peak chromatograms before (in black) and after (in blue) treatment with *N*-hydroxysuccimide (NHS)-activated Sepharose. Peaks in these graphs correspond to putative Glu-C generated peptides of acetylated lysozyme C. *RT*: retention time; NL: normalised intensity level; min: minute.

Before treatment

Protein sequence coverage: 86%

Matched peptides shown in ***bold red***.

```
  1 KVFGRCELAA AMKRHGLDNY RGYSLGNWVC AAKFESNFNT QATNRNTDGS
 51 TDYGILQINS RWWCNDGRTP GSRNLCNIPC SALLSSDITA SVNCAKKIVS
101 DGNGMNAWVA WRNRCKGTDV QAWIRGCRL
```

After treatment

Protein sequence coverage: 18%

Matched peptides shown in ***bold red***.

```
  1 KVFGRCELAA AMKRHGLDNY RGYSLGNWVC AAKFESNFNT QATNRNTDGS
 51 TDYGILQINS RWWCNDGRTP GSRNLCNIPC SALLSSDITA SVNCAKKIVS
101 DGNGMNAWVA WRNRCKGTDV QAWIRGCRL
```

Figure 3.9 Comparison of the sequence coverage of acetylated lysozyme C before and after treatment with *N*-hydroxysuccimide (NHS)-activated Sepharose.

Table 3.5 Comparison of the Glu-C generated peptides of acetylated lysozyme C before and after treatment with NHS-activated Sepharose[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value	Duplicate PSM No.
Before treatment					
KVFGRCE (1 – 7) + Acetyl (N-term); Carbamidomethyl (C)	469.2317	936.4487	32	0.0006	12
KVFGRCE (1 – 7) + Acetyl (N-term); Acetyl (K); Carbamidomethyl (C)	490.2360	978.4593	32	0.00065	23
KVFGRCELAAAMKRHGLD (1 – 18) + Acetyl (N-term); 2 Acetyl (K); Carbamidomethyl (C)	729.0370	2184.0932	30	0.0011	
LAAAMKRHGLD (8 – 18) + Acetyl (K)	408.8887	1223.0444	51	7.7E-6	1
SNFNTQATNRNTD (36 – 48)	741.8313	1481.6495	69	1.3E-7	
SNFNTQATNRNTDGSTD (36 – 52)	921.8875	1841.7776	60	9.2E-7	22
YGILQINSRWWCND (53 – 66) + Carbamidomethyl (C)	912.9256	1823.8413	26	0.0025	
GRTPGSRNLCNIPCSALLSSD (67 – 87) + 2 Carbamidomethyl (C)	759.0336	2274.0845	29	0.0012	1
ITASVNCAKKIVSD (88 – 101) + Acetyl (K); Carbamidomethyl (C)	774.4073	1546.8025	14	0.037	
GNGMNAWVAWRNRCKGTD (102 – 119) + Acetyl (K); Carbamidomethyl (C)	712.3243	2133.9585	24	0.0044	
VQAWIRGCRL (120 – 129) + Carbamidomethyl (C)	629.8446	1257.6764	14	0.037	
After treatment					
KVFGRCE (1 – 7) + Acetyl (N-term); Acetyl (K); Carbamidomethyl (C)	490.2354	978.4593	16	0.023	1
SNFNTQATNRNTD (36 – 48)	741.8319	1481.6495	74	4.2E-8	1
SNFNTQATNRNTDGSTD (36 – 52)	921.8926	1841.7776	65	3.3E-7	22

[a] Orbitrap RAW data were processed and searched in Mascot twice, with cysteine (C) carbamidomethylation set as either a fixed or variable modification. Both Mascot searches produced the same result, which was then edited to show a single PSM each representing one of the 11 significantly identified peptides (*E*-value ≤ 0.05). NHS: *N*-hydroxysuccinimide; *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

3.2.2 Method improvement by refining experimental conditions

The presence of contaminating internal peptides has impeded the use of the NHS-Sepharose approach to selectively identify protein N-termini. To overcome this problem, false-positive peptides such as SNFNTQATNRNTDGSTD of lysozyme C should be further characterised. The outcome of such peptide characterisation may help to fine-tune experimental conditions, thereby leading to a reduced number of false-positive peptides and improved N-terminal enrichment.

Therefore, an investigation was carried out by Hao Zhang (H.Z.) in collaboration with a Final Year Project student, Zimeng Zhang (Z.Z.). This investigation had two primary aims: I. to determine if specific physico-chemical properties are shared by the false-positive peptides; and II. to lower the number of such contaminating peptides by adjusting relevant experiment parameters. As with previous experiments, peptides that resulted from Glu-C digestion of the acetylated lysozyme C were treated with NHS-activated Sepharose for N-terminal enrichment. In addition, bovine serum albumin (BSA, Swiss-Prot ID: ALBU_BOVIN/P02769) was subjected to acetylation and tryptic digestion, and the resulting peptides were treated similarly. H.Z. and Z.Z. jointly performed the experiments in I, and Z.Z. analysed the resulting data. The experiments and data analysis in II were carried out primarily by Z.Z. with help from H.Z.

BSA is also a suitable candidate for this investigation because tryptic digestion of the acetylated BSA produced a 10-residue N-terminal peptide DTHKSEIAHR (with one missed cleavage), which was readily detected by LC-M/MS (Table 3.6). The tandem mass spectrum

Table 3.6 Acetylated N-terminal peptide of BSA (amino acid sequence: DTHKSEIAHR), as identified by a Mascot database search[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value	Duplicate PSM No.
DTHKSEIAHR (1 – 10)	398.5389	1192.5949	48	1.7E-5	1
DTHKSEIAHR (1 – 10) + Acetyl (N-term)	412.5425	1234.6054	59	1.2E-6	5
DTHKSEIAHR (1 – 10) + Acetyl (K)	412.5417	1234.6054	58	1.5E-6	5
DTHKSEIAHR (1 – 10) + Acetyl (N-term); Acetyl (K)	639.3154	1276.6160	63	5.2E-7	19

[a] A single PSM is shown for each significantly identified peptide (*E*-value ≤ 0.05). *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

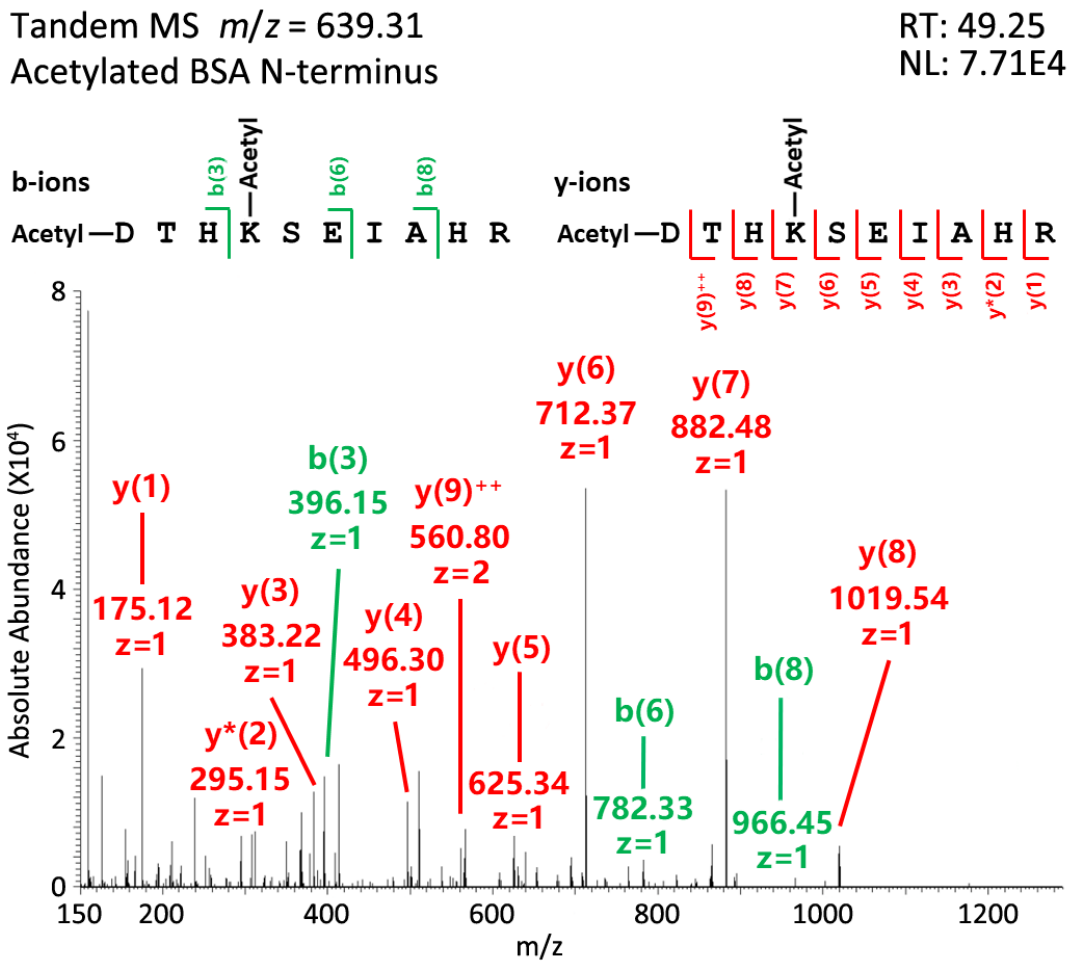


Figure 3.10 Tandem mass spectrum of the precursor ion at *m*/*z* = 639.31 (*z* = 2), which corresponds to the acetylated N-terminal peptide of BSA (amino acid sequence: DTHKSEIAHR). The spectrum was matched to this peptide based on the *in silico* prediction of fragment ions (see Table 3.7). *RT*: retention time; NL: normalised intensity level; *m*/*z*: mass-to-charge ratio.

Table 3.7 Fragment ions (*in silico* predicted) of the peptide DTHKSEIAHR. Fragment ions matched to the experimental data are shown in red.

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	158.0448	79.5260			140.0342	70.5207	D							10
2	259.0925	130.0499			241.0819	121.0446	T	1120.5858	560.7965	1103.5592	552.2833	1102.5752	551.7912	9
3	396.1514	198.5793			378.1408	189.5740	H	*1019.5381*	510.2727	1002.5116	501.7594	1001.5275	501.2674	8
4	566.2569	283.6321	549.2304	275.1188	548.2463	274.6268	K	*882.4792*	441.7432	865.4526	433.2300	864.4686	432.7380	7
5	653.2889	327.1481	636.2624	318.6348	635.2784	318.1428	S	*712.3737*	356.6905	695.3471	348.1772	694.3631	347.6852	6
6	782.3315	391.6694	765.3050	383.1561	764.3210	382.6641	E	*625.3416*	313.1745	608.3151	304.6612	607.3311	304.1692	5
7	895.4156	448.2114	878.3890	439.6982	877.4050	439.2061	I	*496.2990*	248.6532	479.2725	240.1399			4
8	966.4527	483.7300	949.4262	475.2167	948.4421	474.7247	A	*383.2150*	192.1111	366.1884	183.5979			3
9	1103.5116	552.2594	1086.4851	543.7462	1085.5010	543.2542	H	312.1779	156.5926	295.1513	148.0793			2
10							R	*175.1190*	88.0631	158.0924	79.5498			1

for this peptide shows that the *m*/*z* of nine *y*-ions were matched to the *in silico* predicted values. The *b*-series fragment ions are derived from the amino parts of a peptide and hence their *m*/*z* values are directly affected by N-terminal modifications. The nature of this modification is then determined according to the mass shift calculated by comparing the mass of each *b*-ion between a peptide and its modified counterpart. In this case, three *b*-ions were matched: the b_3 ion gained a mass shift of +42.01 Da, whereas both the b_6 and b_8 ions gained a mass shift of +84.02 Da. Taken together, these results indicate the addition of two acetyl groups, one at the N-terminus and the other on the K residue at the fourth position.

As described in Chapter 2 (Materials and Methods), lysozyme C or BSA was acetylated, proteolysed (Glu-C for lysozyme C and trypsin for BSA), and treated with NHS-activated Sepharose for N-terminal enrichment. The purpose of choosing different proteases for different proteins was to generate N-terminal peptides appropriate for MS analysis in terms of sequence length. To identify possible means to further remove internal peptides, seven experiment parameters were manipulated (e.g. the use of different coupling reagents or chaotropic agents). At the data analysis stage, the sequence coverage results before and after treatment with NHS-activated Sepharose were employed to calculate a normalised efficiency of internal peptide removal (Equation 3.1). The normalised efficiency in each parameter alteration was then compared to determine the most appropriate strategy to improve the NHS-Sepharose approach. Details in each parameter alternation are described together with their respective results.

$$\text{Normalised Efficiency} = \left(\frac{SC_{before} - SC_{after}}{SC_{before} - SC_{N-peptide}}\right) \times 100\%$$

Equation 3.1 Equation to calculate **Normalised Efficiency** of removing internal peptides. $\boldsymbol{SC_{before}}$ represents the sequence coverage before treatment with *N*-hydroxysuccinimide (NHS)-activated Sepharose (control sample); $\boldsymbol{SC_{after}}$ is the sequence coverage after the treatment (treated sample); $\boldsymbol{SC_{N\text{-}peptide}}$ corresponds to the percentage of the N-terminal peptide in the entire protein sequence (i.e. the ideal percentage that a perfect result would achieve with no false-positives).

In the case of removing internal peptides of acetylated lysozyme C, the previously identified internal peptide SNFNTQATNRNTDGSTD was still present after treatment with NHS-activated Sepharose. On the other hand, the acetylated N-terminal peptide of BSA was identified together with multiple internal peptides after the treatment (Table 3.8). Several conjectures were therefore put forward to explain this observation. First, such false-positive peptides were produced in higher abundance than others (e.g. due to favourable cleavage sites), so that these peptides were more frequently detected by MS. Second, these false-positive

peptides share certain common physico-chemical features, which prevented the efficient removal by NHS-activated Sepharose (e.g. non-covalent binding to N-terminal peptides). Third, these peptides were more efficiently ionised than other peptides hence were more readily detectable in a mass spectrometer.

Table 3.8 Contaminating tryptic peptides of acetylated BSA, as identified by a Mascot database search after treatment with NHS-activated Sepharose[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value	Duplicate PSM No.
CASIQKFGER (199 – 208) + Acetyl (K); Carbamidomethyl (C)	619.3031	1236.5921	26	0.0025	
ALKAWSVAR (209 – 217) + Acetyl (K)	522.3036	1042.5923	55	3.4E-6	2
VHKECCHGDLLECADDR (240 – 256) + Acetyl (K); 3 Carbamidomethyl (C)	719.3020	2154.8881	48	1.5E-5	1
RHPEYAVSVLLR (336 – 347)	720.4081	1438.8045	82	6.0E-9	26
LAKEYEATLEECCAK (348 – 362) + Acetyl (K); 2 Carbamidomethyl (C)	928.9211	1855.8331	20	0.01	
KVPQVSTPTLVEVSR (413 – 427)	820.4692	1638.9305	30	0.001	1
KVPQVSTPTLVEVSR (413 – 427) + Acetyl (K)	841.4739	1680.9410	16	0.023	
CCTKPESER (436 – 444) + Acetyl (K); 2 Carbamidomethyl (C)	604.7554	1207.4961	42	6.5E-5	
LCVLHEKTPVSEK (459 – 471) + Acetyl (K); Carbamidomethyl (C)	527.9473	1580.8232	40	0.00011	1
VTKCCTESLVNR (472 – 483) + Acetyl (K); 2 Carbamidomethyl (C)	754.8639	1507.7123	55	3.4E-6	
CCTESLVNR (475 – 483) + 2 Carbamidomethyl (C)	569.7515	1137.4907	19	0.013	

[a] Orbitrap RAW data were processed and searched in Mascot twice, with cysteine (C) carbamidomethylation set as either a fixed or variable modification. Both Mascot searches produced the same result, which was then edited to show a single PSM for each of the 11 internal tryptic peptides identified with high confidence (*E*-value ≤ 0.05). NHS: *N*-hydroxysuccinimide; *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

These conjectures were first addressed by comparing the signal intensity (expressed as normalised intensity level, NL) between contaminating (i.e. false-positive) and other proteolytic peptides in the RAW data of control samples (without treatment with NHS-

activated Sepharose). In quantitative proteomics, NL measurement is a crude yet convenient method to approximate peptide abundance as compared to isotopic labelling approaches. The signal intensity of a given peptide ion is correlated with its abundance despite sometimes questionable accuracy. However, there was no distinct difference in the abundance between contaminating and other proteolytic peptides according to the results of NL comparison.

Next, the hydrophobicity of the first two amino acid residues in each contaminating peptide was evaluated in an attempt to identify common physico-chemical properties shared by these peptides. The hydrophobicity of peptide N-terminus was measured by counting the average number of hydrophobic residues at the first two positions of each peptide. Often in protein post-translational modifications (PTMs) and artificial chemical modifications, the reactivity of the target amino acid residue is critically influenced by the following residue. For instance, methionine aminopeptidases (MetAPs) specifically excise the N-terminal methionine (M) residue when it is followed by an alanine (A), C, G, proline (P), serine (S), threonine (T), or V residue (Bonissone *et al.*, 2013). On the other hand, a P residue at the second position disfavours the tryptic digestion at K/R residues (Olsen *et al.*, 2004) or the Rapoport's salt (RS)-mediated transamination at the N-terminus of a peptide (Witus *et al.*, 2013). Therefore, it was hypothesised in the present study that both the N-terminus and the second residue might influence the removal of contaminating peptides.

As shown in Figure 3.11, there is a trend of increased hydrophobicity at the N-termini of contaminating peptides compared to other peptides despite the statistical insignificance (*P*-value > 0.05 as calculated by a Student's *t*-test). In a protein, the hydrophobicity of amino acid residues contributes to its structure formation. For instance, hydrophobic residues often reside in the middle of integral membrane proteins, forming lipid-interacting domains. Therefore, the presence of hydrophobic residues at the N-termini of these contaminating peptides might prevent interaction with NHS-activated Sepharose by burying the free α-amino groups. Alternatively, the hydrophobic N-termini of these peptides might associate with each other through hydrophobic interactions. Again, this would sterically hinder access by NHS-activated Sepharose.

Following the determination of common properties shared by contaminating peptides, a range of experiment parameters were altered individually to explore the means to remove internal peptides more efficiently. The altered parameters include: the amount of acetylation reagents, the type of coupling reagents, the amount of coupling reagents, coupling duration, reaction pH, the number of coupling reactions, and the use of chaotropic agents.

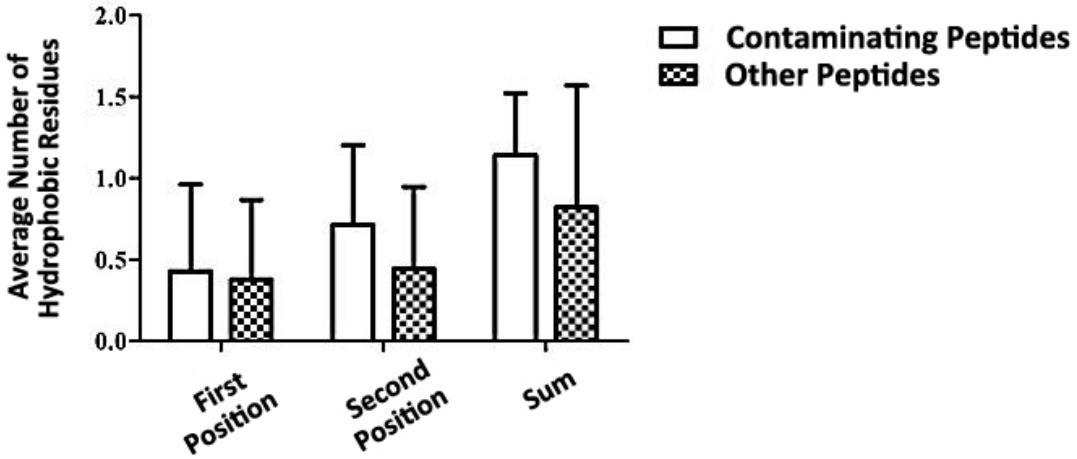


Figure 3.11 Average number of hydrophobic amino acid residues at the N-termini of contaminating or other proteolytic peptides (the first amino acid position, the second position, or the sum of both). In each group, a Student's *t*-test was performed between the contaminating and other peptides with no statistical significance reported (*P*-value > 0.05). Error bars indicate standard deviation.

The amount of sulfo-NHS acetate (i.e. the acetylation reagent) was the first parameter explored in this study. The original strategy employed 1 mg of the acetylation reagent to block free amines at protein N-termini. However, the N-termini of several internal peptides were also acetylated according to LC-MS/MS results. This observation led to the conjecture that the amount of acetylation reagents was excessive in the original protocol, thereby resulting in "carry-over" and hence the acetylation of internal peptides after protease digestion. Accordingly, the amount of sulfo-NHS acetate was decreased from 1 mg to 100 µg. In addition, the "carry-over" problem may be solved by ensuring thorough removal of the acetylation reagent prior to protease digestion. As a result, the amount of the quenching reagent (Tris(2-aminoethyl)amine, polymer-bound) was concomitantly increased from 5 mg to 10 mg. The coupling reaction between NHS-activated Sepharose and proteolytic peptides was not modified.

As reflected by the protein sequence coverage, internal peptides were present in both lysozyme C and BSA samples when the original NHS-Sepharose protocol was followed (Figure 3.12A & B). Interestingly, the experiments on lysozyme C and BSA produced different results with respect to the effect of reducing the amount of acetylation reagents: the normalised efficiency of internal peptide removal was elevated in lysozyme C samples, but the efficiency remained largely constant in BSA samples (Figure 3.12C).

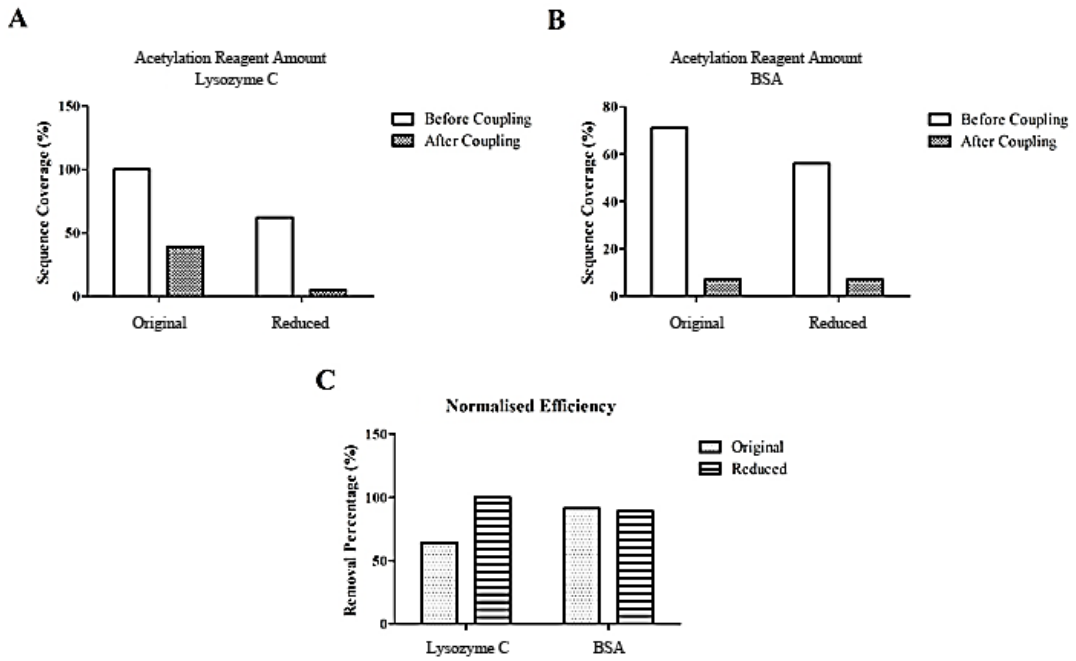


Figure 3.12 Effect of the amount of acetylation reagent on the removal of internal peptides of lysozyme C (**A**) or BSA (**B**). (**C**) Normalised coupling efficiency calculated for both proteins.

The discrepancy in these results may be a product of multiple factors. At the reduced amount, the acetylation reagent might no longer interfere with the subsequent coupling step, as reflected by the increased coupling efficiency in lysozyme C samples. This result agrees with our conjecture. On the other hand, reducing the amount of acetylation reagents might lead to less acetylation of BSA on K side chains. As trypsin digests proteins at the C-terminal side of K and R residues when not followed by a P residue, less acetylation at K residues could result in the yield of more tryptic peptides in BSA samples. The amount of tryptic peptides of BSA exceeded the coupling capacity of NHS-activated Sepharose, so that internal peptide removal was not improved in BSA samples.

Second, two different types of amine-reactive beads (NHS-activated Sepharose and NHS-activated magnetic beads) were employed in parallel to remove internal peptides of lysozyme C or BSA after digestion with their respective proteases. The two beads were then compared in terms of the coupling efficiency. This study aimed to determine whether this coupling reagent is a limiting factor in this approach. As the name implies, the second coupling reagent also possesses an NHS ester group that specifically conjugates with a primary amine to form a stable amide bond and releases the NHS group. However, the reactive group is on the surface of magnetic beads that enable convenient bead separation by use of a strong neodymium magnet instead of centrifugation. The experiment with the magnetic beads was performed exactly the same as the original NHS-Sepharose protocol, apart from replacing NHS-activated Sepharose with the magnetic beads.

As shown in Figure 3.13, the sequence coverage of lysozyme C was lowered to a similar extent by both NHS-activated Sepharose and the magnetic beads. In contrast, treatment with the magnetic beads did not lead to a decrease in the sequence coverage of BSA. When converted to the normalised efficiency of removing internal peptides, these results clearly showed that NHS-activated Sepharose outperformed the magnetic beads. Therefore, NHS-activated Sepharose was chosen as the coupling reagent in the following experiments.

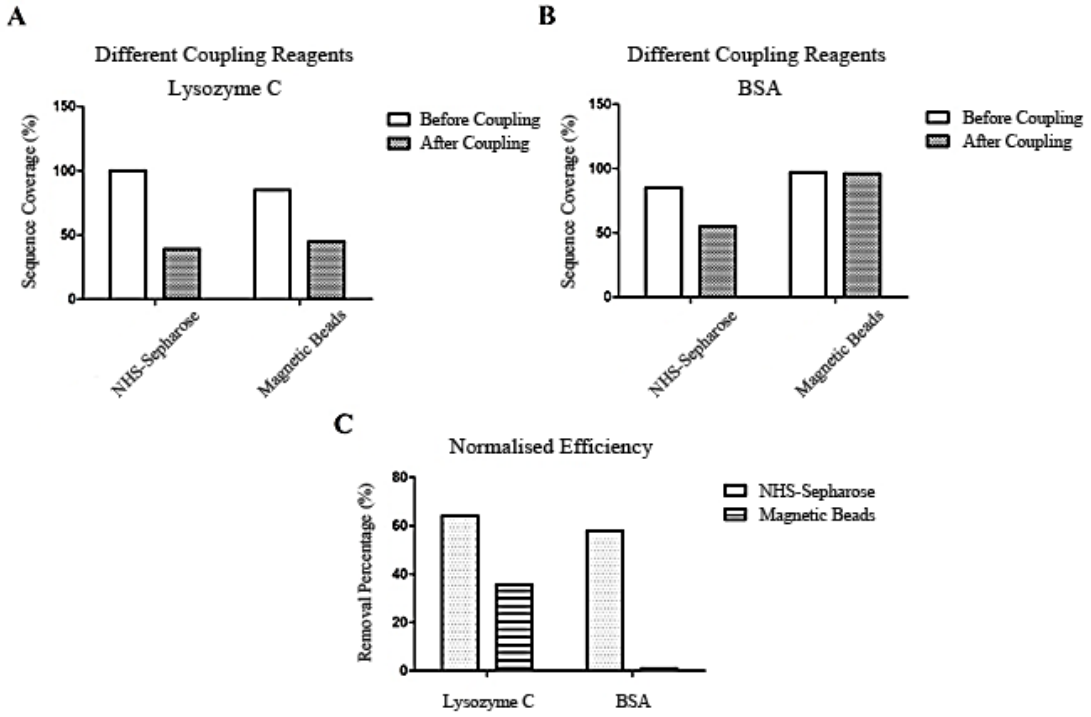


Figure 3.13 Efficacy of two different coupling reagents (NHS-activated Sepharose or magnetic beads) for removing the internal peptides of lysozyme C (**A**) or BSA (**B**), reflected as the changes in sequence coverage after coupling. (**C**) Normalised efficiency of removing internal peptides calculated for the two beads. NHS: *N*-hydroxysuccinimide.

The binding capacity of these beads is determined by the concentration of reactive NHS groups on bead surface. In addition, their binding capacity may also be affected by the spacer arm that bridges the beads and the NHS groups. The spacer arm has been shown to help overcome steric hindrance (Cuatrecasas *et al.*, 1968) and to influence electrostatic and non-electrostatic interactions (DePhillips *et al.*, 2004). However, the manufacturers of these two beads have not disclosed information regarding the concentration of reactive groups or the length of spacer arm. Such information could help researchers to choose or design coupling reagents with a higher binding capacity.

The third explored experiment parameter was the amount of the coupling reagent (NHS-activated Sepharose), based on the hypothesis that the abundance of proteolytic peptides

might exceed the binding capacity of the coupling reagent. Proteolytic peptides of lysozyme C was incubated with different volumes of the coupling reagent at 0, 100, 200, or 300 µl in four parallel experiments. The hypothesis was that use of the largest amount of the coupling reagent would result in the lowest number of internal peptides. Figure 3.14 shows that the normalised efficiency did increase in direct proportion to the increased amount of the coupling reagent. Among all four experiment groups, the largest volume of coupling reagent (300 µl) exhibited the highest binding capacity, in agreement with the hypothesis. Therefore, the increase of the amount of coupling reagent improves the removal of internal peptides by the NHS-Sepharose approach. Further increase (beyond 300 µl) was not tested in this experiment because the reagent consumption would be significant. However, it may afford further improvements to the NHS-Sepharose approach.

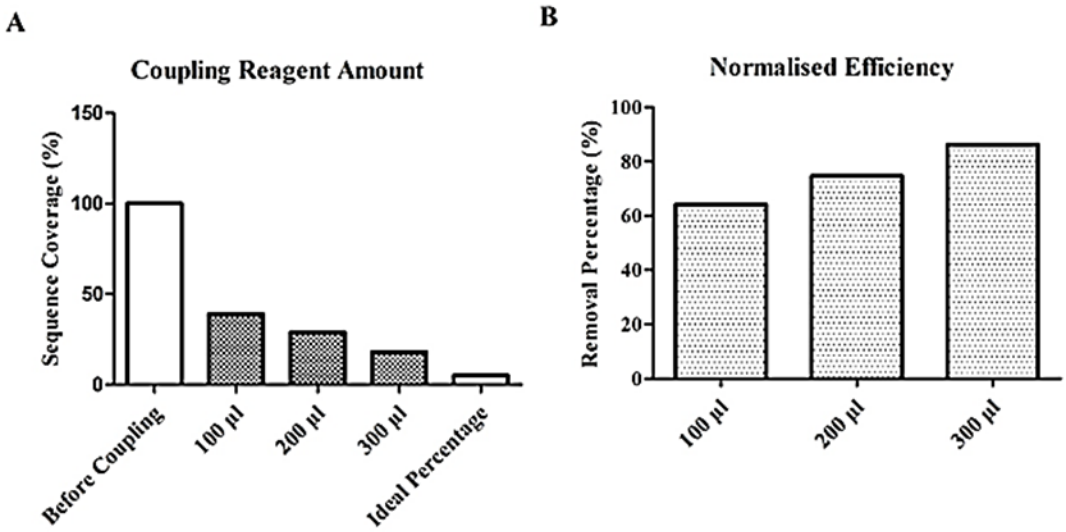


Figure 3.14 (**A**) Effect of the amount of coupling reagent (NHS-activated Sepharose) on the removal of internal peptides of lysozyme C. (**B**) Normalised coupling efficiency calculated for each group. NHS: *N*-hydroxysuccinimide.

The next explored experiment parameter was the time of incubation with the coupling reagent, which potentially affected the outcome of removing internal peptides. In this experiment, tryptic peptides of BSA were incubated with 100 µl of the coupling reagent at either room temperature (RT) or 4 °C for varying amounts of time. The RT group was set at 1, 2, 3, 4, or 8 h; whereas the 4 °C group was at 3, 6, 9, 12, or 24 h. Upon conversion from the sequence coverage data, no linear correlation was observed between the normalised efficiency of internal peptide removal and the incubation time (Figure 3.15). However, there was a deduced trend of increase in the normalised efficiency from 1-h to 4-h in the RT group and from 3-h to 12-h in the 4 °C group, respectively. The normalised efficiency then dropped from 4-h to 8-h in the RT group and from 12-h to 24-h in the 4 °C group, respectively.

Possible explanations for this fluctuation may reside in the underlying mechanism of the reaction between NHS esters and primary amines. As explained previously, NHS-activated Sepharose couples with internal peptides through a conjugation reaction between NHS esters and primary amines at the N-termini of these peptides, forming stable amide bonds while releasing NHS (Figure 3.16A). Simultaneously, the NHS groups are self-decomposing through a hydrolysis reaction (Figure 3.16B). Water molecules (in this case from phosphate buffered saline, PBS) compete with primary amines at peptide N-termini for the reactive NHS groups. With extended incubation time, a larger fraction of NHS esters were hydrolysed instead of conjugating with primary amines, leading to a decreased coupling efficiency. Alternatively, the stable amide bonds or the spacer arm may break down during incubation, releasing the coupled internal peptides.

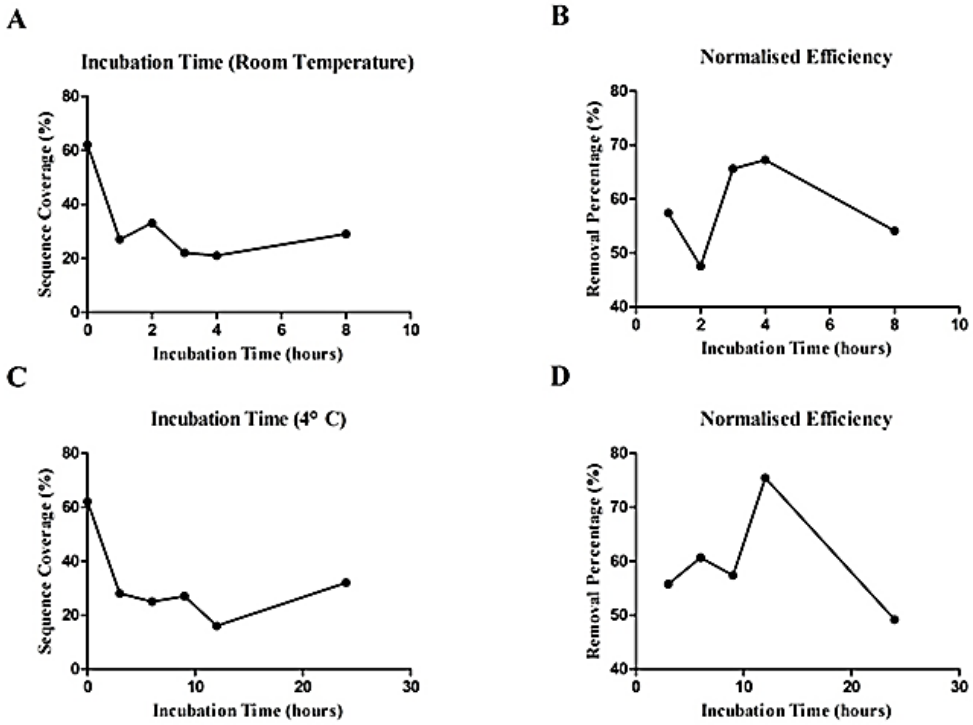


Figure 3.15 Effect of incubation time at room temperature (RT, **A**) or 4 °C (**C**) on the removal of internal peptides of BSA. Normalised coupling efficiency calculated for each group of incubation time at RT (**B**) or 4 °C (**D**).

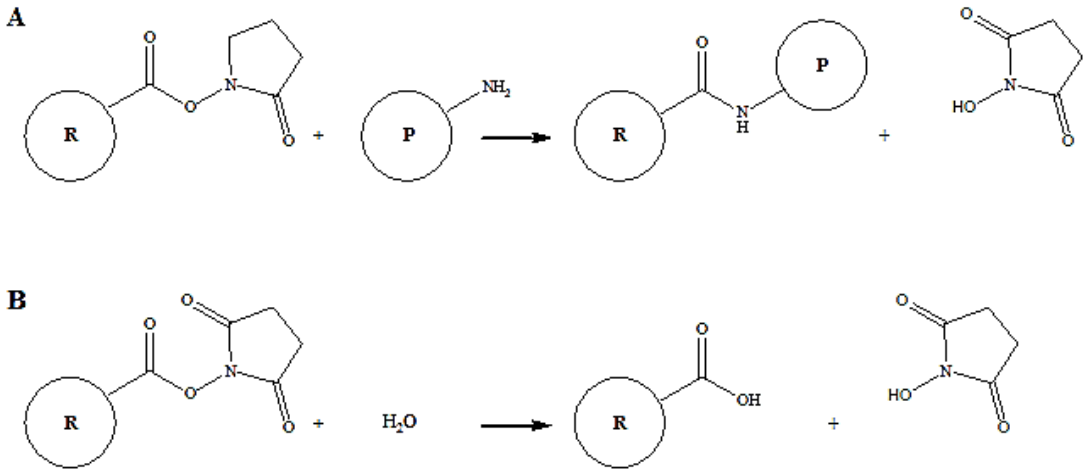


Figure 3.16 Illustrated reaction mechanism of the conjugation between primary amines and *N*-hydroxysuccinimide (NHS) esters (**A**) or the hydrolysis of NHS esters (**B**). Ⓡ standards for reagent whereas Ⓟ standards for protein.

The pH value is critical to the coupling reaction between NHS-activated Sepharose and free α-amino groups at peptide N-termini. To improve the NHS-Sepharose approach, this experiment parameter was explored in this study as well. The NHS-amine reaction was usually performed at neutral to slight basic conditions. In the original protocol, the coupling step takes place in an amine-free buffer (i.e. PBS) at pH 7.5. Four parallel experiments were performed where lysozyme C or BSA was acetylated and proteolysed without coupling with NHS-activated Sepharose, or incubated with the beads at pH 7.5, 8, or 8.5. This study aimed to evaluate the differences in the normalised coupling efficiency among the three pH values.

For lysozyme C, a steady decrease in sequence coverage was observed when the pH was elevated from 7.5 to 8.5. The difference in the sequence coverage was statistically significant between pH 7.5 and 8.5 (*P*-value = 0.0084 as determined by one-way analysis of variance, ANOVA). Therefore, the normalised efficiency appeared to correlate positively with the pH of the coupling reaction. This correlation was less evident for BSA, but the normalised efficiency at pH 8 or 8.5 was still higher than that at pH 7.5 (Figure 3.17). The increase in pH might enhance the coupling reaction possibly through maintaining a structure of peptides that confers higher accessibility by NHS esters. Alternatively, higher pH might elevate the reactivity of primary amines as nucleophiles. However, it needs to be emphasised that much higher pH can adversely affect the NHS-amine conjugation since the hydrolysis of NHS esters dominates over conjugation under very basic conditions.

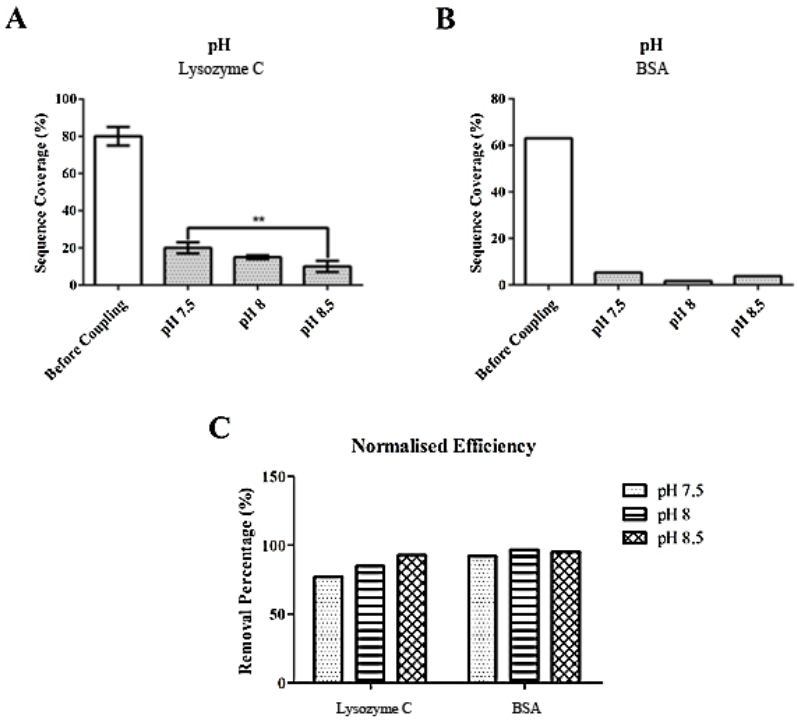


Figure 3.17 Effect of coupling reaction pH on the removal of internal peptides of lysozyme C (**A**) or BSA (**B**). (**C**) Normalised coupling efficiency calculated for both proteins. The statistical analysis in **A** was performed using one-way analysis of variance (ANOVA) with the Bonferroni correction (the significance level was set at 0.05). Error bars indicate standard deviation; ** indicates *P*-value ≤ 0.01.

Based on the previous finding that contaminating peptides tend to have more hydrophobic N-termini, the next step in this study was to investigate the influence of chaotropic agents on removing internal peptides by NHS-activated Sepharose. Chaotropic agents such as urea or guanidine hydrochloride (HCl) are compounds that disrupt non-covalent interactions, including hydrogen bonds that maintain protein structures and hydrophobic interactions between proteins (Makhatadze and Privalov, 1992, Lim *et al.*, 2009). Therefore, chaotropic agents are typically employed to unfold proteins and solubilise hydrophobic proteins. In the present study, high concentrations of urea and guanidine HCl were added to the existing NHS-Sepharose protocol based on the hypothesis that these chaotropic agents would improve the removal of contaminating peptides. Lysozyme C and BSA were incubated with 8 M urea or 6 M guanidine HCl for 1 h prior to the coupling with NHS-activated Sepharose. For both proteins, the removal of internal peptides without adding chaotropic agents served as a control group.

The addition of urea appeared to further reduce the sequence coverage of lysozyme C after the treatment with NHS-activated Sepharose, although no statistical significance was reached (*P*-value > 0.05 as calculated by one-way ANOVA). The same experiment was repeated with BSA and produced a similar result, leading to the conclusion that the addition

of urea consistently increased the normalised coupling efficiency for both proteins. In contrast, guanidine HCl was unable to exert the same effect, with only a marginal improvement of the normalised efficiency for BSA. Therefore, urea served as a superior chaotropic agent over guanidine HCl in removing contaminating peptides of lysozyme C and BSA. This conclusion conflicts with previous findings that guanidine HCl is generally more effective than urea in inducing protein unfolding and denaturation (Myers *et al.*, 1995, Lim *et al.*, 2009). Nevertheless, the removal of contaminating peptides was improved by chaotropic agents, which supports the previous hypothesis that the contaminating peptides may physically associate with protein N-termini through hydrophobic interactions or that the N-termini of these peptides are inaccessible to the coupling reagent due to their relative hydrophobicity.

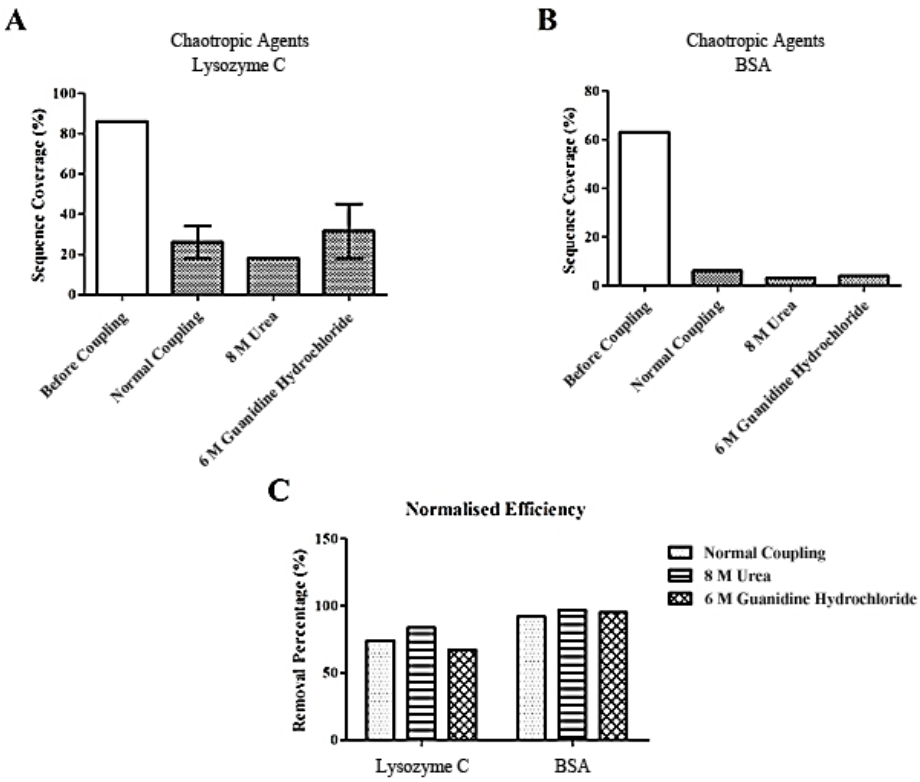


Figure 3.18 Effect of different chaotropic agents (8 M urea or 6 M guanidine hydrochloride) on the removal of internal peptides of lysozyme C (**A**) or BSA (**B**). (**C**) Normalised efficiency calculated for both proteins. The statistical analysis in **A** was performed using one-way analysis of variance (ANOVA) with Bonferroni correction (the significance level was set at 0.05). Error bars indicate standard deviation.

The alteration of the number of coupling reactions was the last experiment parameter explored in this systematic study. In the original protocol, proteolytic peptides of lysozyme C or BSA were treated with two batches of NHS-activated Sepharose: the first at RT for 4 h and the second at 4 °C overnight. This number was increased in the present study to 3 or 4, and

the amount of NHS-activated Sepharose was also doubled in each coupling step. The hypothesis was that an increase in the number of coupling reactions would further reduce the number of contaminating peptides till the saturation of free α-amino groups on internal peptides with the NHS groups on the beads.

As shown in Figure 3.19, the removal of internal peptides was highly efficient as a result of more repeats of coupling. The normalised coupling efficiency was already at 100 % for lysozyme C and ~ 90 % for BSA with two repeats of coupling, probably due to the doubled amount of the coupling reagent in this investigation. For BSA, an even higher efficiency (> 90 %) was achieved after four repeats of coupling. Consequently, it is concluded that an increase in the number of coupling reactions positively affected the removal of contaminating peptides. However, a higher coupling efficiency was achieved at the cost of more handling time and reagent consumption. A balance between performance and feasibility is thus required in order to promote the use of this approach.

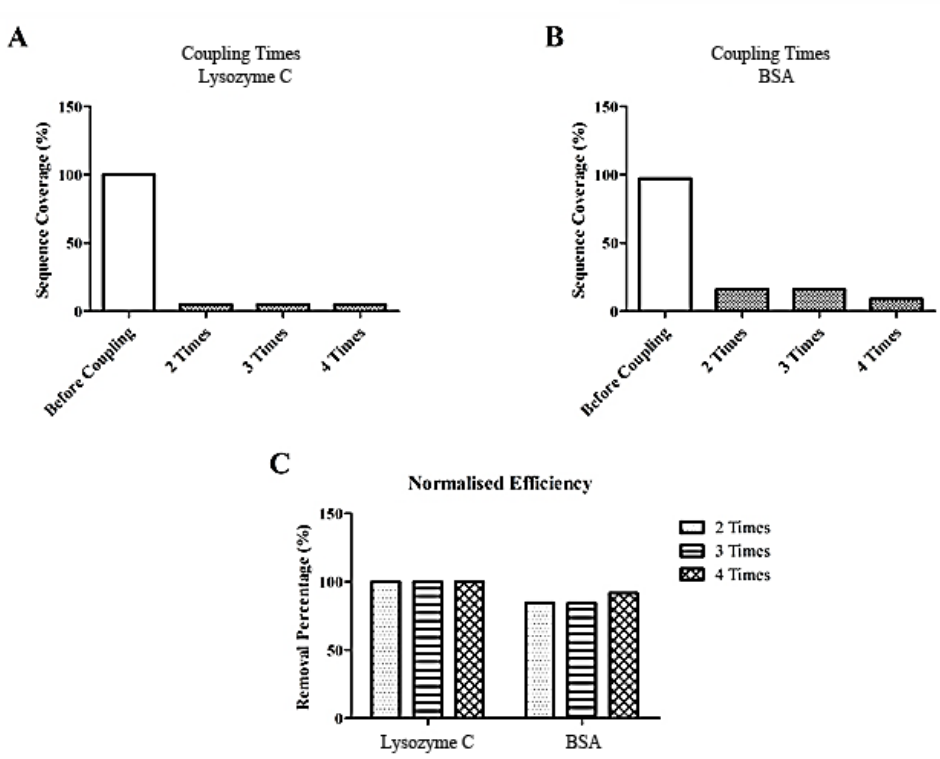


Figure 3.19 Effect of the number of coupling reactions on the removal of internal peptides of lysozyme C (**A**) or BSA (**B**). (**C**) Normalised coupling efficiency calculated for both proteins.

In light of the outcome of altering seven experiment parameters, an integrated method was developed in order to remove internal peptides of BSA with the highest efficiency. The protocol of this integrated method is as follows: 100 µg of BSA was acetylated with a reduced amount of sulfo-NHS acetate (100 µg) and the reaction was quenched with an increased

amount of quenching reagents (10 mg). After tryptic digestion, 50 µg of tryptic peptides were incubated with either 8 M urea or 6 M guanidine HCl for 1 h, or directly treated with 4 x 200 µl of NHS-activated Sepharose at pH 8.5: two 4-h incubations at RT and two 12-h incubations at 4 °C. The sample without incubation with chaotropic agents served as a "normal coupling" control. The other 50 µg of BSA peptides served as a "before coupling" control, i.e. no treatment with NHS-activated Sepharose. In addition, the data acquired using the original NHS-Sepharose method (see Table 3.8) were also included for comparison.

Without the use of chaotropic agents, this integrated method was shown to efficiently remove the internal peptides of BSA. As shown in Figure 3.20, the normalised efficiency of this method is 20 % higher than that using the original protocol. The addition of either urea or guanidine HCl could further increase the normalised efficiency to 100 %. In conclusion, this integrated method removed all of the contaminating peptides for BSA, and could be tested on other model proteins or a complex mixture of proteins.

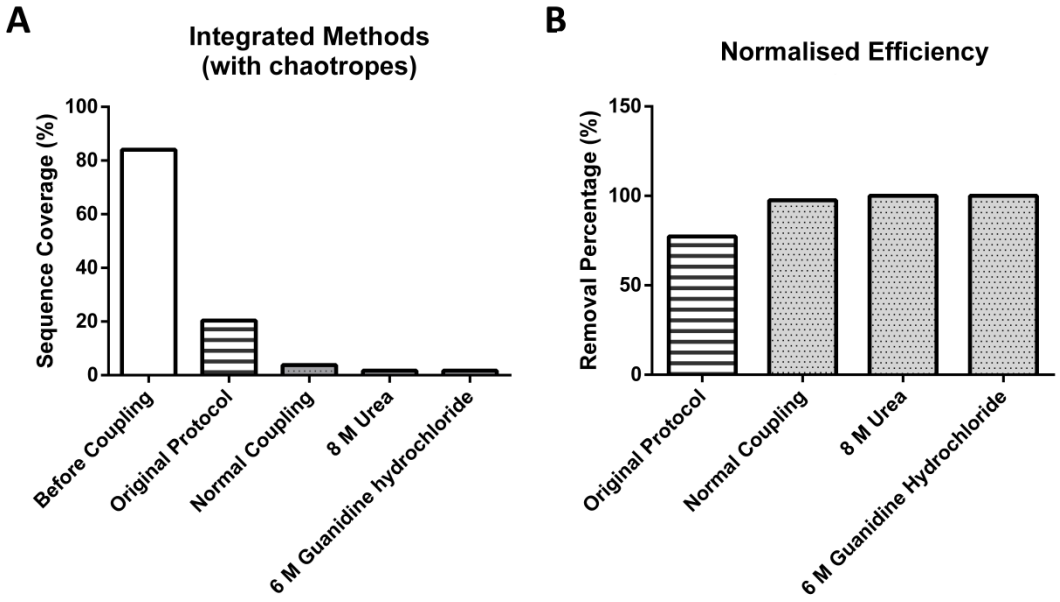


Figure 3.20 (**A**) Outcome of the integrated NHS-Sepharose methods (with or without chaotropes), reflected as the changes in sequence coverage of BSA after coupling. The integrated methods incorporated alterations in the following experiment parameters: a reduction in the amount of acetylation reagents, an increase in the amount of quenching reagents, a 1-h incubation with chaotropic agents (either 8 M urea or 6 M guanidine hydrochloride), an increase in the amount of coupling reagents, and a doubling of the coupling times (see section 3.2.2). The "before coupling" control shows the sequence coverage of BSA before internal peptide removal. The "original protocol" group refers to implementation of the original NHS-Sepharose method described by McDonald and Beynon (2006). The "normal coupling" control corresponds to the use of the integrated method without chaotropic agents. (**B**) Normalised efficiency calculated for each method.

3.2.3 Use of SATA and CA as an alternative approach to select protein N-termini

As discussed in Chapter 1, the N-termini of proteins possess highly informative biological data, as they indicate the protein coding region of a gene, protein half-lives, and sometimes biological functions of these proteins (Varshavsky, 2011, Varland *et al.*, 2015). The importance of protein N-termini is partly attributed to the PTMs that take place at protein N-termini, including N-terminal methionine excision (NME), Nt-myristoylation, and Nt-acetylation (reviewed in Giglione *et al.*, 2015). As a result, analysis of N-terminal PTMs is an integral part of proteomic research. On the other hand, free protein N-termini (without any PTM) are of equal biological importance and thus require comprehensive analysis. As described previously, the original NHS-Sepharose approach has been tested on individual model proteins and improved by systematic refinement of experimental conditions. This approach combines two chemical reactions, amine acetylation and NHS-amine conjugation, to enrich for N-terminal peptides with high efficiency (for lysozyme C and BSA). However, it does not distinguish free N-termini with the acetylated ones due to the use of sulfo-NHS acetate.

The present section describes the work to expand the use of the NHS-Sepharose approach to identify protein N-termini by replacing sulfo-NHS acetate with alternative amine-reactive chemicals. Two such reagents, *N*-succinimidyl *S*-acetylthioacetate (SATA) and citraconic anhydride (CA), were selected for method development as they form stable amide bonds with primary amines. In principle, such chemical modifications are also compatible with internal peptide removal by NHS-activated Sepharose. The reaction mechanisms of SATA and CA are shown in Figure 3.21. These changes to the original protocol should enable the identification of both endogenous Nt-acetylation and free protein N-termini and their discrimination according to the attached chemical groups.

As mentioned previously, the protocol of the NHS-Sepharose approach included the acetylation of intact proteins and protease digestion of the acetylated proteins, followed by coupling with NHS-activated Sepharose to remove internal peptides with free N-termini. This protocol was modified by blocking protein α- and ε-amino groups with either SATA or CA instead of sulfo-NHS acetate. In addition, individual model proteins (lysozyme C and BSA) were replaced by a complex mixture of proteins extracted from Jurkat T-lymphocytes. Details of Jurkat cell culture, protein extraction, and protein quantitation are described in Chapter 2.

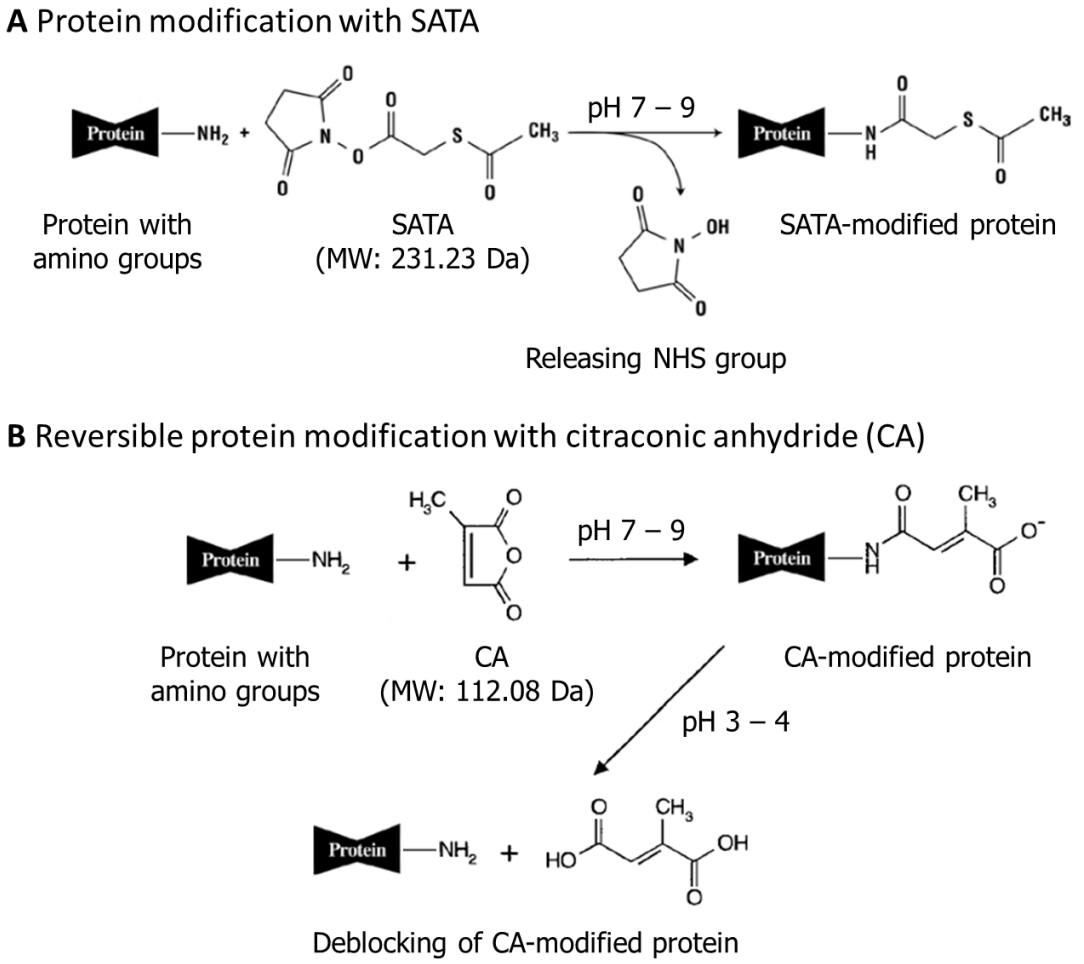


Figure 3.21 Illustrated mechanism of the reaction between *N*-succinimidyl *S*-acetylthioacetate (SATA; **A**) or citraconic anhydride (CA; **B**) and primary amines (modified from Hermanson, 1996).

The first objective of this study was to obtain a comprehensive list of proteins extracted from Jurkat T-cells by shotgun proteomics. Without the reaction with SATA/CA or internal peptide removal by NHS-activated Sepharose, Jurkat proteins were denatured, reduced, alkylated, and digested with trypsin as described in Chapter 2. The tryptic peptides were then separated on a 3-meter monolithic column and analysed by MS. Using Mascot software package, the recorded RAW data were converted to a peak list, which was then searched against a decoy database (Swiss-Prot, taxonomy = *Homo sapiens*). The search parameters were set as follows: trypsin with maximally two missed cleavages, carbamidomethylation (C) as fixed modification, oxidation (M) and acetylation (Protein N-term/K) as variable modifications. The search results were further processed in R to generate a non-duplicate protein list. In total, 2,253 proteins were unambiguously identified from three biological replicates of Jurkat protein samples (N = 3; ProteomeXchange dataset ID: PXD009340). Table 3.9 shows an example of the proteins identified by the Mascot database search.

Table 3.9 Representative proteins identified in Jurkat T-cell extracts by a Mascot database search[a].

Protein Name	Swiss-Prot ID	Peptide No.[b]
(E3-independent) E2 ubiquitin-conjugating enzyme	UBE2O_HUMAN (Q9C0C9)	4
1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase	MTND_HUMAN (Q9BV57)	5
10 kDa heat shock protein, mitochondrial	CH10_HUMAN (P61604)	37
14-3-3 protein epsilon	1433E_HUMAN (P62258)	50
14-3-3 protein eta	1433F_HUMAN (Q04917)	26
14-3-3 protein gamma	1433G_HUMAN (P61981)	37

[a] Orbitrap RAW data were processed and searched against the decoy Swiss-Prot database (taxonomy = *Homo sapiens*). [b] Peptide No.: number of tryptic peptides identified with high confidence (*E*-value ≤ 0.05) for each protein; *E*-value: peptide expectation value.

In the next step, we aimed to investigate the feasibility of blocking free protein N-termini and K residues with SATA or CA. In principle, SATA and CA are similar to sulfo-NHS acetate in reacting with primary amines at both protein N-termini and K side chains. In addition, the unique chemical groups added to protein N-termini can be exploited to distinguish free N-termini from endogenously acetylated ones. As described in Chapter 2, Jurkat proteins were treated with either SATA or CA, and a control sample was prepared by omitting the reaction step. In the case of SATA modification, 8 M urea solution was added to the modified proteins to dissolve pellets that resulted from protein precipitation. In contrast, no protein precipitation was observed after CA modification. Excess reactants were removed from the protein samples by ultrafiltration upon the completion of these reactions. The modified proteins were then reduced, denatured, alkylated, and digested with trypsin.

The tryptic peptide samples were quickly analysed on 10 % NuPAGE™ Bis-Tris gels to confirm proteolysis. Details of sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) are described in Chapter 2. Before tryptic digestion, both the SATA- and CA-modified Jurkat protein samples yielded a range of protein bands on the polyacrylamide gel, which were absent in the samples of tryptic peptides as their MW were below the lower detection limit of SDS-PAGE (Figure 3.22). The tryptic peptide samples only yielded a single protein band that corresponds to trypsin (23.3 kDa) based on MW estimation. In conclusion, tryptic digestion was highly efficient and proceeded to completion.

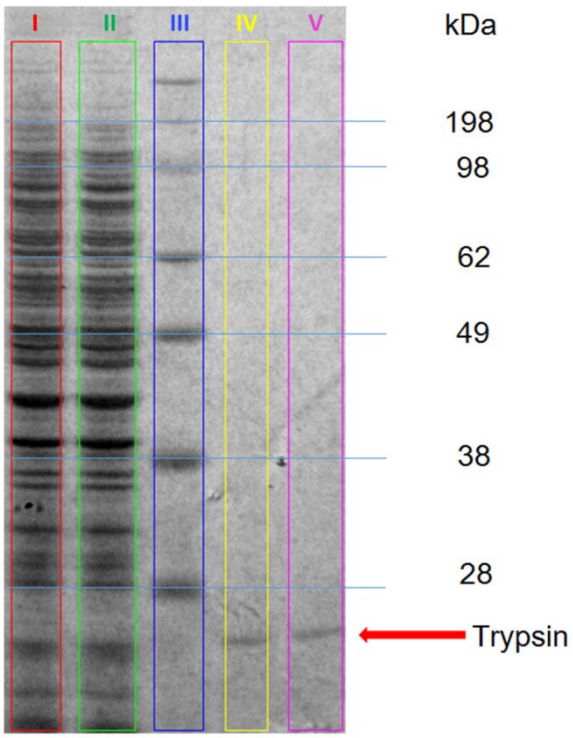


Figure 3.22 SDS-PAGE analysis of Jurkat proteins before and after tryptic digestion. Four different samples were separated on a 10 % NuPAGE™ Bis-Tris gel together with a protein molecular weight (MW) marker: I. SATA-modified proteins without tryptic digestion; II. CA-modified proteins without tryptic digestion; III. protein MW marker; IV. tryptic digested sample of the SATA-modified proteins; V. tryptic digested sample of CA-modified proteins. A single band in both lane IV and V corresponds to trypsin (red arrow). SDS-PAGE: sodium dodecyl sulfate–polyacrylamide gel electrophoresis; SATA: *N*-succinimidyl *S*-acetylthioacetate; CA: citraconic anhydride.

Subsequently, the tryptic peptides from SA- or CA-modified Jurkat proteins were treated with NHS-activated Sepharose to remove internal peptides and to enrich for protein N-terminal peptides. The enriched peptides were then subjected to LC-MS/MS analysis for spectral acquisition, and the resulting data were processed and searched against a decoy Swiss-Prot database (taxonomy = *Homo sapiens*). For data analysis of SATA-treated samples, SATA (Protein N-term/K) was included in the variable modifications in addition to oxidation (M) and acetylation (Protein N-term/K). In contrast, CA (Protein N-term/K) was included in the variable modifications for CA-treated samples.

The SATA modification experiment involved 12 biological replicates (N = 12), and the 12 resulting datasets were pooled to yield a total of 18,779 peptide-spectrum matches (PSMs). Overall, the 12 replicates allowed the identification of 2,801 tryptic peptides, 11 % of which (312) had been modified with SATA. Figure 3.23 shows the cumulative addition of unique tryptic peptides by each replicate. A 12-set Venn diagram visualising all the overlaps between every two replicates could not be drawn due to technical limitations. Nevertheless, our effort

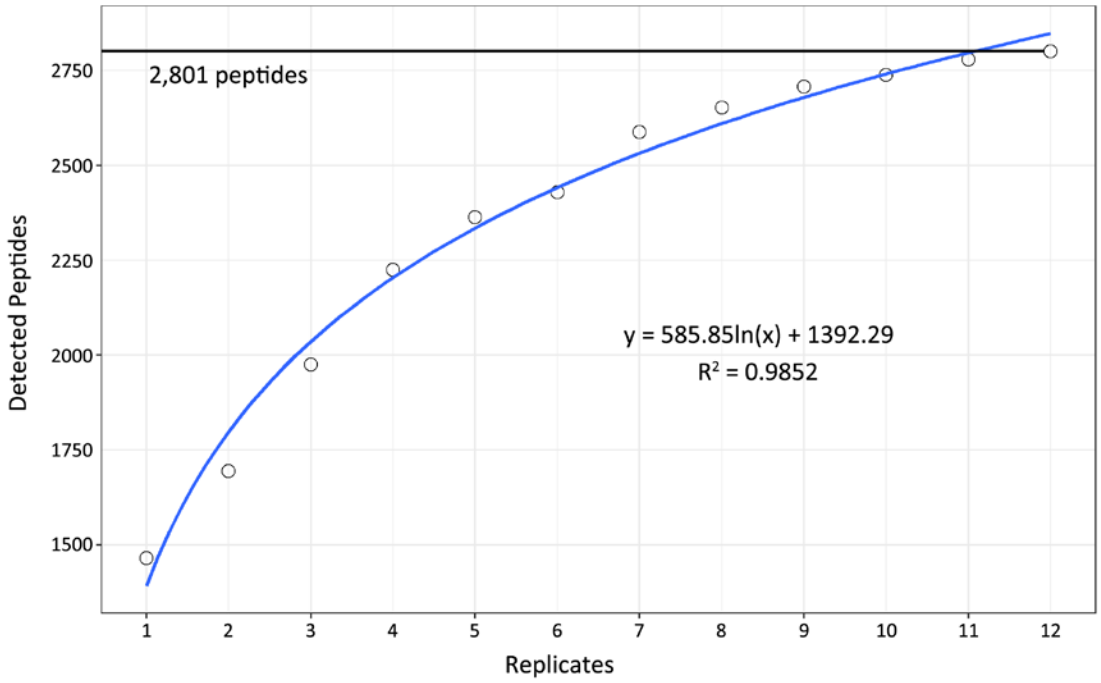


Figure 3.23 Saturation curve of unique tryptic peptides detected by replicate analyses in the *N*-succinimidyl *S*-acetylthioacetate (SATA) modification experiment. A logarithmic curve (blue) was fitted to the experimental data using R, and the total peptide number is reflected by the black straight line.

to produce a Venn diagram for six replicates covering > 85 % of the total peptides can be found in Appendix 1. The 2,801 peptides in turn could be assigned to 397 Jurkat proteins under the "two-peptide rule" (i.e. a protein must be identified on the basis of ≥ 2 significant peptide hits). Importantly, 106 of the peptides were assigned as protein N-termini to 83 proteins. These protein N-termini were further sorted in three different forms: Nt-acetylated, SATA-modified, and free. As shown in Table 3.10, 88 out of the 106 peptides were Nt-acetylated, 4 were SATA-modified, and 14 were free of N-terminal PTMs.

Table 3.10 Distribution of 106 peptides assigned as protein N-termini in the SATA modification experiment. The 106 peptides are divided into three groups: Nt-acetylated, SATA-modified, and free.

N-terminal PTMs	Nt-acetyl	SATA	Free	Total
Assigned peptides	88	4	14	106
Identified proteins	74	3	12	83 (89)[a]

[a] The number in parentheses indicates the sum of protein numbers from all three groups. It is higher than the true number of proteins with assigned N-termini since a single protein N-terminus could be identified in different PTM states. PTM: post-translational modification; SATA: *N*-succinimidyl *S*-acetylthioacetate; Nt-acetyl: N-terminal acetylation.

Table 3.11 gives a list of SATA-modified N-terminal peptides as identified by the Mascot database search. Three proteins were deemed to possess a protected sulfhydryl group (mass shift = +115.99 Da) at the N-terminus as a result of SATA modification. For example, the SATA-modified peptide AGELADKKDR was assigned as the N-terminus of DDX23_HUMAN, a probable ATP-dependent RNA helicase. The tandem spectrum for this peptide is shown in Figure 3.24, where three *b*-ions and seven *y*-ions were identified by matching their *m*/*z* to the *in silico* predicted values (Table 3.12). The mass of each *y*-ion remained unchanged after SATA modification, whilst all the *b*-ions gained the mass shift corresponding to SATA modification.

In addition, eukaryotic initiation factor 4A-III (Swiss-Prot ID: IF4A3_HUMAN) was shown to exist in two proteoforms with different N-termini (MATTATMATSGSAR or ATTATMATSGSARK). The co-existence of both N-terminal peptides was validated by further cross-referencing with the N-terminal annotation of IF4A3_HUMAN in the online Swiss-Prot database (Van Damme *et al.*, 2012). In conclusion, SATA could be employed as an amine-reactive chemical to modify and thereby identify free N-termini of proteins such as DDX23_HUMAN and IF4A3_HUMAN.

Table 3.11 SATA-modified N-terminal peptides as identified by a Mascot database search[a].

Protein ID (Swiss-Prot)	Peptide-Spectrum Match (Start – End)	*m*/*z*[b]	MW	Score	*E*-value
DDX23_HUMAN (Q9BUQ8)	AGELADKKDR (2 – 11) + SATA (N-term)	609.80	1217.57	30	0.049
EF2_HUMAN (P13639)	VNFTVDQIR (2 – 10) + SATA (N-term)	604.30	1206.57	31	0.0013
IF4A3_HUMAN (P38919)	MATTATMATSGSAR (1 – 14) + SATA (N-term); 2 Oxidation (M)	752.8095	1503.6004	23	0.015
	ATTATMATSGSARK (2 – 15) + SATA (N-term); Oxidation (M)	743.3491	1484.6599	29	0.039

[a] A single PSM is shown for each significantly identified peptide (*E*-value ≤ 0.05). [b] SATA: *N*-succinimidyl *S*-acetylthioacetate; *m*/*z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; PSM: peptide-spectrum match.

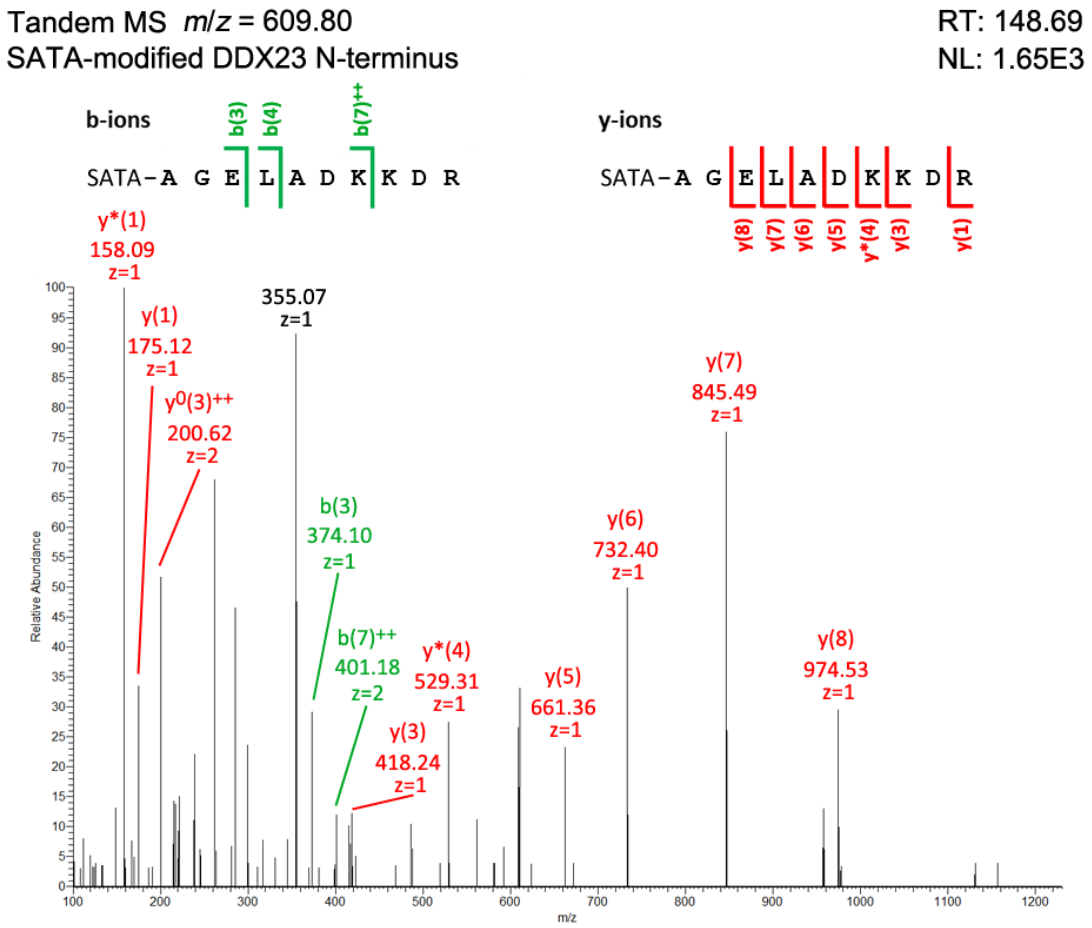


Figure 3.24 Tandem mass spectrum of the SATA-modified N-terminal peptide of probable ATP-dependent RNA helicase DDX23 (amino acid sequence: AGELADKKDR). The detected *b*- and *y*-ions are shown in green and red, respectively. SATA: *N*-succinimidyl *S*-acetylthioacetate; *RT*: retention time; NL: normalised intensity level.

Table 3.12 Fragment ions (*in silico* predicted) of the *N*-succinimidyl *S*-acetylthioacetate (SATA)-modified peptide AGELADKKDR. Fragment ions matched to the experimental data are shown in red.

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	188.0376	94.5224					A							10
2	245.0591	123.0332					G	1031.5480	516.2776	1014.5215	507.7644	1013.5374	507.2724	9
3	374.1016	187.5545			356.0911	178.5492	E	***974.5265***	487.7669	957.5000	479.2536	956.5160	478.7616	8
4	487.1857	244.0965			469.1751	235.0912	L	***845.4839***	423.2456	828.4574	414.7323	827.4734	414.2403	7
5	558.2228	279.6151			540.2123	270.6098	A	***732.3999***	366.7036	715.3733	358.1903	714.3893	357.6983	6
6	673.2498	337.1285			655.2392	328.1232	D	***661.3628***	331.1850	644.3362	322.6717	643.3522	322.1797	5
7	801.3447	401.1760	784.3182	392.6627	783.3342	392.1707	K	546.3358	273.6715	529.3093	265.1583	528.3253	264.6663	4
8	929.4397	465.2235	912.4131	456.7102	911.4291	456.2182	K	***418.2409***	209.6241	401.2143	201.1108	400.2303	200.6188	3
9	1044.4666	522.7370	1027.4401	514.2237	1026.4561	513.7317	D	290.1459	145.5766	273.1193	137.0633	272.1353	136.5713	2
10							R	***175.1190***	88.0631	158.0924	79.5498			1

Based on the results from the SATA modification experiment, 21 % of the proteins were assigned with N-termini (83/397). The vast majority of the protein N-termini were Nt-acetylated (83 %, 88/106), whereas the SATA-modified ones only constituted 4 % of the total set (4/106). In principle, all the free protein N-termini should be blocked with SATA after the reaction, otherwise they would be removed by NHS-activated Sepharose. Contrary to this assumption, 13 % of the identified N-terminal peptides were neither acetylated nor modified with SATA (i.e. free N-termini). The number of free N-termini was 2.5-fold higher than that of the SATA-modified N-termini. These observations suggested that SATA modification was not as efficient as chemical acetylation, and that NHS-activated Sepharose did not fully remove peptides with free N-termini.

The same data analysis procedure was repeated on the Jurkat proteins treated with CA. A total of 12 CA-treated samples were analysed by LC-MS/MS and the resulting data were pooled to generate a list that contained 31,979 PSMs. Overall, 3,611 tryptic peptides (including 42 CA-modified ones) or 570 Jurkat proteins were identified from the combined list of PSMs. The cumulative addition of unique peptides by each replicate is shown in Figure 3.25, which suggests that saturated peptide detection was achieved by 12 replicate analyses. As with the SATA data analysis, a Venn diagram was also drawn to visualise shared peptides for six replicate analyses covering > 95 % of the total peptides (see Appendix 1).

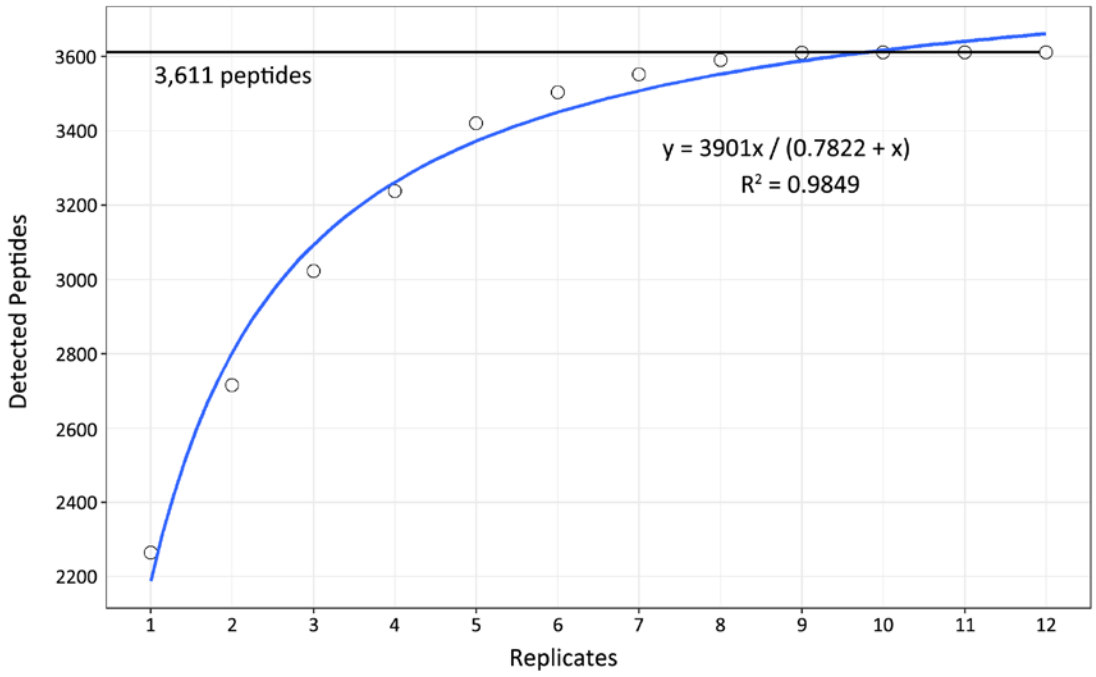


Figure 3.25 Saturation curve of unique tryptic peptides detected by replicate analyses in the citraconic anhydride (CA) modification experiment. The curve in blue was fitted through a nonlinear regression analysis in R, and the total peptide number is reflected by the straight line in black.

Among the detected tryptic peptides, 311 were assigned as protein N-termini and such peptides were derived from 143 Jurkat proteins (Table 3.13). The identified N-terminal peptides were also divided into three groups according to N-terminal modifications: Nt-acetylated, CA-modified, and free. With a number of 240, acetylated protein N-termini accounted for 77 % of all the identified N-terminal peptides; the rest 23 % were entirely free protein N-termini (71). It should be emphasised that N-terminal modification with CA could not be confirmed in this experiment: the efficiency of CA modification might be lower than that of SATA modification, or CA modification might have been reversed before LC-MS/MS analysis. As mentioned previously, the amide bond formed by CA modification may dissociate under acidic conditions, i.e. pH 3 – 4 (Dixon and Perham, 1968). Since tryptic peptides from the CA-treated samples were exposed to 0.5 % (v/v) formic acid solutions at the desalting step, this might have reversed the amide bond formation.

Table 3.13 Distribution of 311 peptides assigned as protein N-termini in the CA modification experiment. The 311 peptides are divided into three groups: Nt-acetylated, CA-modified, and free.

N-terminal PTMs	Nt-acetyl	CA	Free	Total
Assigned peptides	**240**	**0**	**71**	**311**
Identified proteins	**117**	**0**	**36**	**143 (153)[a]**

[a] The number in parentheses indicates the sum of protein numbers from all three groups. It is higher than the true number of proteins with assigned N-termini since a single protein N-terminus could be identified in different PTM states. CA: citraconic anhydride; PTM: post-translational modification; Nt-acetyl: N-terminal acetylation.

Collectively, data from SATA- and CA-treated samples allowed a preliminary survey of different types of protein N-termini in Jurkat T-cells, despite the incomplete blocking by SATA or CA. All the data from both the SATA and CA experiments were merged to assign the identified protein N-termini to four different groups: SATA-modified, CA-modified, Nt-acetylated, or free (Table 3.14). Overall, the N-termini of 229 Jurkat proteins were assigned, corresponding to 10 % of the total proteins identified in the shotgun experiment (229/2,253, Figure 3.26A). Nt-acetylated proteins accounted for 82 % of all the proteins with assigned N-termini (188/229). In contrast, the N-termini of < 1 % of such proteins (1/229) were modified with SATA. In addition, free N-termini were directly assigned to 17 % of these proteins (40/229) without the aid of N-terminal modification with SATA/CA (Figure 3.26B). The complete list of proteins with an assigned N-terminus is shown in Appendix 2.

As SATA- and CA-modified N-terminal peptides were also derived from free protein N-termini, these three groups together contributed to 18 % of all the Jurkat proteins with assigned N-termini (41/229). The ratio between the acetylated and free N-termini (82 % versus 18 %) is consistent with the estimation from other studies: 80 – 90 % of human cytosolic proteins are endogenously acetylated at the N-terminus (Kalvik and Arnesen, 2013, Giglione *et al.*, 2015). However, the present survey on N-terminal modifications is incomplete as only 10 % of the identified proteins were assigned with N-termini. The present approach thus requires further improvements on the efficiency of both SATA/CA modification and peptide removal. A refined strategy can then be applied to validate and expand the current results of Jurkat N-terminalomic analysis.

Table 3.14 Statistics of assigned N-termini of Jurkat proteins, which are classified by their PTM states[a].

N-terminal PTMs	SATA	CA	Nt-acetyl	Free	Total
Assigned peptides	**4**	**0**	**270**	**79**	**353**
Identified proteins	**1 (3)**	**0**	**188**	**40 (54)**	**229 (245)[b]**

[a] Protein N-termini are divided into four groups: SATA-modified, CA-modified, Nt-acetylated, and free.
[b] The number in parentheses indicates the sum of protein numbers from all four groups. It is higher than the true number of proteins with assigned N-termini since a single protein N-terminus could be identified in different PTM states. PTM: post-translational modification; SATA: *N*-succinimidyl *S*-acetylthioacetate; CA: citraconic anhydride; Nt-acetyl: N-terminal acetylation.

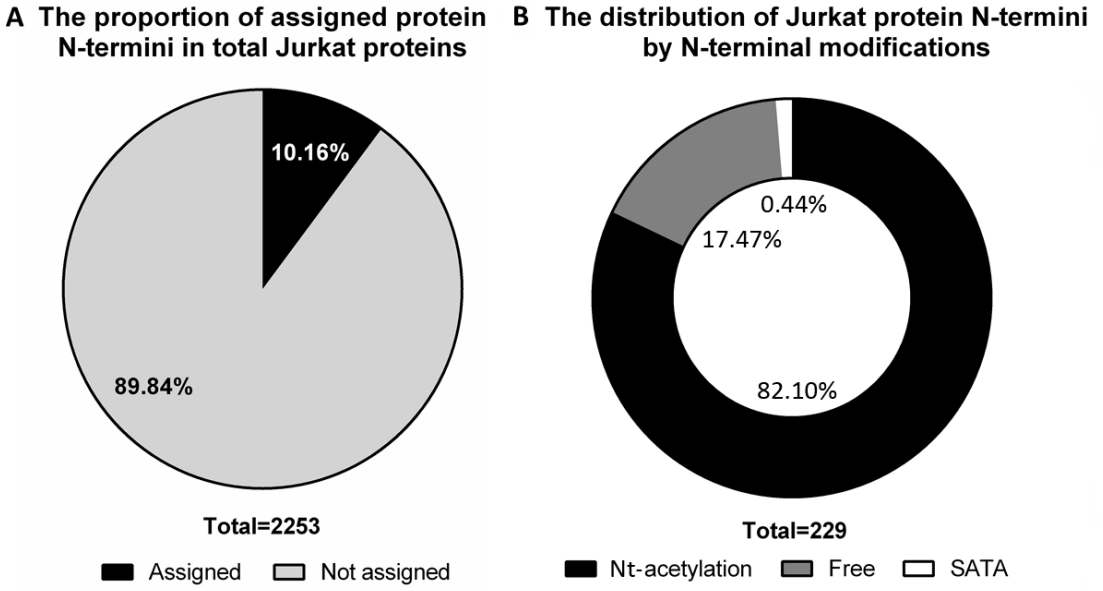


Figure 3.26 Statistics of Jurkat proteins with assigned N-termini. **(A)** The proportion of assigned protein N-termini (229) in the surveyed Jurkat proteome (2,253). **(B)** The distribution of the assigned protein N-termini with respect to N-terminal modifications: SATA (1), free (40), and Nt-acetylation (188). No CA-modified protein N-termini were detected. SATA: *N*-succinimidyl *S*-acetylthioacetate; CA: citraconic anhydride.

3.3 Discussion

In positional proteomics, negative selection strategies are employed to remove interfering peptides in order to retain peptides with defined positions. In principle, such techniques are therefore ideal for the global identification of protein N-termini. Such proteomic studies will potentially detect discrepancies between genome-derived protein N-termini and experimental data due to alternative splicing, alternative translation initiation, PTMs, etc. Since top-down proteomics faces numerous technical challenges regarding MS instrumentation and data processing, proteolysis is still required for standard proteomic analysis (Catherman *et al.*, 2014, Chen *et al.*, 2018). As a result, authentic protein N-termini are overwhelmed by a vast mixture of peptides produced by protease digestion. These newly formed peptides are often termed as internal peptides or *neo*-peptides. One particular feature of these internal peptides is that they possess a free α-amino group at the newly formed N-terminus.

As one of the most elegant negative selection strategies, the NHS-Sepharose approach enriches the N-terminal peptides of proteins by combining the acetylation of all protein amino groups before proteolysis and the removal of internal peptides by NHS-activated Sepharose afterwards. We intend to fully exploit the potential of this approach to study protein N-termini and further expand the use of this approach in studying protease substrates. The present study comprised three phases where the NHS-Sepharose approach was implemented on different protein samples with a gradually elevating complexity. The first phase was to select the N-terminal peptide of a well-characterised protein, lysozyme C, by strictly following the original protocol by McDonald and Beynon (2006). This experiment was an ideal entry point and also served to acquire the essential skills in LC-MS/MS and data analysis.

In this experiment, lysozyme C was chosen from six model protein candidates including horse myoglobin, horse cytochrome C, BSA, yeast enolase, and rabbit phosphorylase b. The choice was made on the basis of two major facts. First, this protein commences with a free N-terminus, whereas several other candidates are N-terminally blocked by endogenous acetylation. Only proteins with free N-termini will be useful in assessing the efficiency of the reaction with sulfo-NHS acetate. Second, proteolysis of lysozyme C with Glu-C yielded a mixture of proteolytic peptides (including a 7-residue N-terminal peptide, KVFGRCE) that could be completely identified by LC-MS/MS, corresponding to the 100 % sequence coverage. Complete sequence coverage was not achieved with other protein candidates.

The NHS-Sepharose approach consists of three major steps: intact protein blocking by acetylation, protease digestion, and removal of internal peptides by NHS-activated Sepharose. It was demonstrated through LC-MS/MS analysis that acetylation took place as expected at the N-terminus of lysozyme C. Two acetyl groups (MW: 42.01 Da each) were attached to the N-terminal K residue due to the detected mass shift of +84.02 Da. Acetyl groups were also added to K residues in other internal peptides according to LC-MS/MS results. All the above results suggest that primary amine acetylation is an efficient chemical reaction to block protein N-termini and K side chains. The KK motif of the peptide ITAVNCAKKIVSD is the only example of incomplete acetylation. Peptides with a KK motif have been shown to exhibit the lowest kinetic constant for tryptic digestion (Slechtova *et al.*, 2015). Three models were thus put forward to explain the inefficient tryptic digestion of such peptides: the lack of trypsin exopeptidase activity, the competition for the enzyme active site, and less favourable ionic interactions. The observation that the second K residue also inhibited K acetylation supports the latter two models.

The 86 % sequence coverage also suggests that protease digestion with Glu-C proceeded to a satisfactory level. The efficiency of internal peptide removal was directly evaluated by comparing the sequence coverage of lysozyme C before and after the coupling with NHS-activate Sepharose. The decrease in the sequence coverage from 86 % to 18 % suggests that the vast majority of Glu-C generated peptides were removed by the coupling reagent. The 18 % sequence coverage corresponds to two peptides: the acetylated protein N-terminus (KVFGRCE) and an internal peptide (SNFNTQATNRNTDGSTD). In the present study, this internal peptide can be easily distinguished from the authentic protein N-terminus. However, the presence of internal peptides would undoubtedly interfere with the identification of true protein N-termini in an entirely different context, i.e. the analysis of unknown proteins or proteins without validated N-termini.

The internal peptide SNFNTQATNRNTDGSTD was repeatedly and consistently identified in lysozyme C samples after the coupling with NHS-activated Sepharose. The three-dimensional (3D) structure of lysozyme C was therefore inspected to identify possible causes for the persistence of this peptide. One conjecture is that the internal peptide is physically associated with the protein N-terminus. However, the 3D structure analysis suggests that the peptide is highly accessible to external agents and should be removed by NHS-activated Sepharose after proteolysis (Figure 3.27). The analysis also shows that the D48 residue within the internal peptide is located in a loop structure and hence might not be cut efficiently by Glu-C. Consequently, the internal peptide contains one missed cleavage.

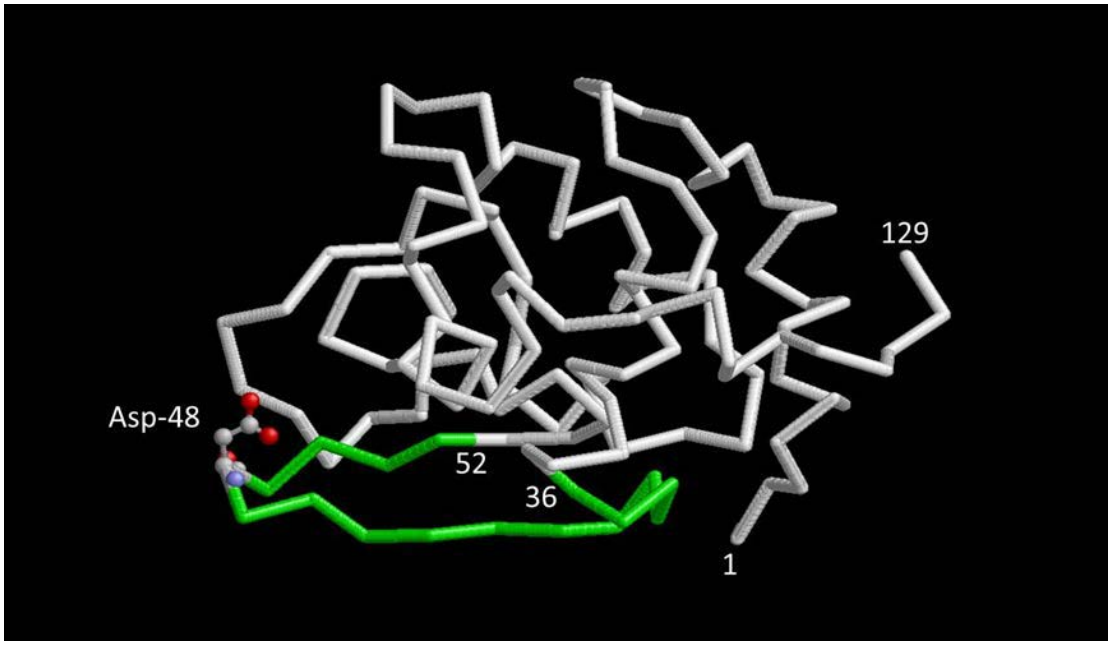


Figure 3.27 Schematic representation of the 3D structure of lysozyme C showing the location of peptide SNFNTQATNRNTDGSTD (36 – 52) in green, and highlighting the position of D48 in a loop within the peptide. The structure 4B0D (Cipriani *et al.*, 2012) was retrieved from the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) and visualised using the programme RasWin.

The second phase of this project was conducted in response to the detection of internal peptides after the negative selection of the N-terminal peptide of lysozyme C. In this experiment, BSA was digested with trypsin to serve as a comparator for testing the NHS-Sepharose approach. Similar to lysozyme C, the results with BSA samples also suggest that the coupling reagent did not efficiently remove internal peptides. Instead of only one contaminating peptide from lysozyme C samples, several internal peptides of BSA remained in the aqueous phase after the coupling step. However, these contaminating peptides together offered an opportunity to study physico-chemical properties that are shared in common by these peptides. Although the sample size is small, it appears that the hydrophobicity of peptide N-termini (defined here as the first two amino acid residues) is associated with the persistence of these contaminating peptides in the samples of enriched protein N-termini. This finding guided subsequent studies to reduce the number of such peptides, with an emphasis on hydrophobicity-related solutions.

The efforts to improve the NHS-Sepharose approach consist of seven individual changes to the existing protocol: reducing the amount of acetylation reagents, using an alternative coupling reagent (i.e. NHS-activated magnetic beads), increasing the amount of coupling reagents, extending the duration of coupling reactions, increasing the number of coupling reactions, increasing the pH of coupling reactions, and introducing chaotropic agents (e.g. 8

M urea or 6 M guanidine HCl). From the results obtained, some of these changes improved the outcome of coupling with NHS-activated Sepharose, leading to less contaminating peptides. Meanwhile, these changes exerted additive effects and were hence integrated into an optimised protocol (Table 3.15). Ultimately, it was possible to achieve a 100 % efficiency of internal peptide removal using this integrated method, whereas the authentic N-terminal peptide can still be readily detected by LC-MS/MS.

Table 3.15 Comparison between the original and optimised protocols of the NHS-Sepharose approach[a].

Procedures	**Original protocol**	**Optimised protocol**
Primary amine acetylation	2-h reaction with 1 mg sulfo-NHS acetate in 20 mM Na_2CO_3 (pH 8.5)	2-h reaction with 100 µg sulfo-NHS acetate in 20 mM Na_2CO_3 (pH 8.5)
Quenching	1-h incubation with 5 mg Tris(2-aminoethyl)amine, polymer-bound	1-h incubation with 10 mg Tris(2-aminoethyl)amine, polymer-bound
Incubation with chaotropic agents	None	1-h incubation with 8 M urea or 6 M guanidine HCl
Coupling with NHS-activated Sepharose	1 x 4-h incubation at RT and 1 x 12-h at 4 °C, each with 100 µl NHS-activated Sepharose in PBS, pH 7.5	2 x 4-h incubations at RT and 2 x 12-h at 4 °C, each with 200 µl NHS-activated Sepharose in PBS, pH 8.5

[a] Only experimental procedures that are different between the two protocols are listed here. NHS: *N*-hydroxysuccinimide; RT: room temperature; PBS: phosphate-buffered saline.

Despite this success, several limitations can be identified. First, certain results of this study have not been statistically validated. Due to time constraints, several experiments were conducted without biological replicates on the same proteins. This problem was partially compensated by using two different proteins (lysozyme C and BSA) for the same experiment. But it would be inappropriate to directly combine the results from these two proteins for statistical analysis. Future studies should focus on obtaining more biological replicates and employing a larger cohort of proteins for these experiments.

The second limitation is the possibility of false negatives. LC-MS/MS analysis is intrinsically limited by the size of peptides to be identified. For example, peptides of less than five amino acid residues will not be identified by Mascot database searches. On the other hand, the *m*/*z* values of ionised peptides may fall out of the specified *m*/*z* range (150 – 2000) so that the peptides will not be recorded by the instruments. These issues can be partially alleviated by use of different proteases in parallel experiments. Different protease treatments could produce N-terminal peptides with different lengths, some of which may fall in the detection

range. In addition, the integrated method has more steps in the protocol (e.g. the incubation with chaotropic agents, more repeats of incubation with coupling reagents, etc.). These additional steps will lead to an inevitable sample loss and potentially the loss of authentic protein N-termini.

In view of the above limitations, the refined NHS-Sepharose approach should better serve as a complementary strategy to validate the results obtained by other N-terminalomic approaches, such as COFRADIC, N-TAILS, and Subtiligase. Given sufficient time and resources, N-terminalomic studies would benefit from employing multiple approaches in parallel to validate and expand the obtained data.

Due to the prevalence of endogenous Nt-acetylation in eukaryotic proteins, the original protocol by McDonald and Beynon (2006) takes advantage of isotopic chemistry to differentiate free and endogenously acetylated N-termini. However this strategy is impractical due to the expense of the reagents and the inconvenience of using radioisotopes. Without isotopic labelling, free protein N-termini are concealed by the reaction with sulfo-NHS acetate and become indistinguishable from the endogenously acetylated ones. Therefore, the third phase of this project aimed to employ alternative amine-reactive chemistries to uniquely modify free protein N-termini for unambiguous identification. Ideally, these reactions should attach a protein N-terminal tag that is different to the acetyl group, and they should be as efficient as the reaction with sulfo-NHS acetate.

SATA and CA are the two chemicals selected for this study. In principle, these two chemicals are capable of blocking primary amines on both protein N-termini and K side chains. After the modification of primary amines and coupling with NHS-activated Sepharose, proteolytic peptides with a protected sulfhydryl group (derived from SATA modification) or a terminal carboxyl group (derived from CA modification) at the N-terminus should be assigned to their cognate proteins as free N-termini. A complex mixture of proteins were employed for these experiments, which were extracted from Jurkat T-cells (immortalised human T-lymphocytes). Overall, 2,253 human proteins were unambiguously identified in a preliminary survey of the Jurkat proteome. These proteins were further treated with SATA or CA before tryptic digestion and coupling with NHS-activated Sepharose. The experiment results demonstrated that free N-termini and K side chains could be modified with SATA, leading to a mass shift of +115.99 Da to the singly modified peptides.

In total, 10 % of the surveyed Jurkat proteome was assigned with protein N-termini, including 188 Nt-acetylated proteins, 40 proteins with free N-termini, and one protein with N-terminal SATA modification. The ratio between Nt-acetylated proteins and proteins with free N-termini was 82 % to 18 %, which is consistent with the prediction (80 % to 20 %). However, these results also reflect the fact that neither SATA nor CA modification is comparable to chemical acetylation in terms of reaction efficiency. At the peptide level, only 1 % of the assigned protein N-termini were modified with SATA and none with CA. In contrast, a large number of free protein N-termini were assigned after the coupling with NHS-activated Sepharose. This result indicates a deficiency in the removal of peptides that contained reactive primary amines, probably due to the limited coupling capacity of NHS-activated Sepharose. Although the detection of free protein N-termini is undesirable, it did allow the profiling of protein N-terminal modifications in the surveyed Jurkat proteome.

A major issue with the present study is that N-termini were only assigned to a small fraction (10 %) of the identified proteins. The lack of N-terminal information possibly arises from instrumental limitations and sample loss. As mentioned previously, MS only acquires data from peptides with appropriate length, hydrophobicity, and ionisability. Tryptic digestion potentially yields a large number of N-terminal peptides that do not meet these criteria. Use of multiple proteases (e.g. trypsin and Glu-C) in parallel experiments may improve the coverage of Jurkat proteome and N-terminal profiling.

On the other hand, protein precipitation was observed during the reaction with SATA. This is likely caused by the removal of positive charges on both protein N-termini and K residues, which in turn may create charge neutrality on the modified proteins and ultimately protein precipitation. Therefore, a large fraction of SATA-modified proteins may not have been detected by LC-MS/MS as the precipitated proteins would have already been removed by centrifugation. Despite the efforts to minimise protein loss by introducing chaotropic agents (e.g. 8 M urea), the precipitated proteins were not completely dissolved. Future studies may benefit from using more potent chaotropic agents (or even MS-compatible detergents) that could more completely dissolve the precipitated proteins.

The absence of CA-modified protein N-termini may not be solely attributed to the low efficiency of this reaction. As illustrated in Figure 3.21, CA modification is a reversible reaction at low pH. When reviewing the procedure of the NHS-Sepharose approach, a low pH step is identified where formic acid was introduced at 0.5 % (v/v) to facilitate peptide desalting. The terminal carboxylate attached by CA modification may thus be released from protein N-termini and K residues. In order to validate this conjecture, further experiments

need to be devised to modify or omit the desalting step. For instance, proteins can be buffer exchanged into volatile buffers (e.g. 50 mM ammonium bicarbonate) after CA modification to obviate the desalting step without damaging LC columns.

To summarise, the present study was carried out to recover the N-terminal peptides of target proteins using the NHS-Sepharose approach. This study consisted of three phases that each aimed to tackle a specific problem: the initial phase assessed the efficacy of this approach with a single defined protein; the second phase improved the original approach based on the knowledge acquired from the first phase; the third phase sought to apply this approach (with alternative reagents) on a complex system to identify protein N-termini. Taken together, the results from all three phases reveal that acetylation is an efficient chemical reaction to block primary amines at protein N-termini and K residues. As an alternative blocking reaction, SATA modification enables the differentiation of free protein N-termini from the acetylated ones through a unique chemical tag attached to protein N-termini. However, N-terminal modification with another blocking reagent (i.e. CA) could not be achieved under the current experimental conditions. SATA and CA modifications are not equally efficient as acetylation in blocking primary amines.

The most essential component of this approach is NHS-activated Sepharose, the coupling reagent used to remove internal peptides after proteolytic digestion. However, the coupling capacity of this reagent does not meet our expectation, as revealed by the results from all three phases. Initially, a specific internal peptide of proteolysed lysozyme C was identified after the coupling with NHS-activated Sepharose. The performance was then improved by altering several experiment parameters so that the internal peptides of proteolysed lysozyme C and BSA were completely removed by NHS-activated Sepharose. Nonetheless, 38 free protein N-termini were identified from the proteolysed Jurkat proteins after the coupling with NHS-activated Sepharose. Ideally, all internal peptides should be removed by this coupling reagent. These results suggest that the present coupling reagent is a major limiting factor in this approach. Future studies could focus on replacing NHS-activated Sepharose with other immobilised amine-reactive beads, such as HPG-ALD polymer used in the N-TAILS approach (Kleifeld *et al.*, 2010).

The results reported in this chapter clearly demonstrates that the success of a negative selection strategy lies in the complete blocking of both α- and ε-amino groups of proteins and the efficient removal of proteolytic peptides that possess free α-amino groups. Both objectives can be difficult to achieve, as reflected by the scarcity of SATA- or CA-modified protein N-termini and the presence of contaminating peptides after coupling with NHS-

activated Sepharose. In contrast, positive selection strategies directly recover protein N-termini without the need to remove internal peptides and hence should have a simpler experimental procedure. However, developing such strategies is more challenging because of the difficulties in distinguishing α- and ε-amino groups. The next chapter describes the efforts to develop a positive selection strategy based on a chemical reaction specific for protein α-amino groups.

Chapter 4. Feasibility of selective transamination for tagging the N-termini of proteins

4.1 Introduction

In parallel to optimising a "negative selection" strategy that relies on the use of *N*-hydroxysuccinimide (NHS)-activated Sepharose to identify free protein amino (N)-termini (i.e. the "NHS-Sepharose" approach), the feasibility of a "positive selection" strategy was also investigated. In N-terminalomics, positive selection strategies take advantage of the semi-unique chemistry of α-amino groups at protein N-termini for the enrichment of such peptides. These methods employ a variety of reagents to selectively modify the α-amino groups so that an affinity tag can subsequently be attached at this position. After proteolysis, the tagged N-terminal peptides are positively selected for proteomic identification. The advantage of positive selection strategies over their negative selection counterparts is that they obviate the need to remove internal peptides after proteolysis. This step is essential to negative selection strategies but has been identified as the major bottleneck in the NHS-Sepharose approach.

N-CLAP (N-terminalomics by chemical labeling of the α-amine of proteins), *O*-methylisourea, and Subtiligase represent the major approaches in the positive selection branch of N-terminalomics. Similar to negative selection strategies, N-CLAP employs an indiscriminate chemical reaction (i.e. Edman degradation chemistry) to block both α-amino groups at protein N-termini and ε-amino groups on lysine (K) side chains (Xu *et al.*, 2009). Different from negative selection strategies though, N-CLAP then selectively excises the blocked N-terminal residues from the rest of the proteins and thus exposes the α-amino groups at the newly formed protein N-termini. The reactive α-amino group can be exploited to attach an amine-reactive biotin tag (e.g. NHS-SS-biotin) to the *neo*-N-terminus of the shortened protein for affinity purification (AP). On the other hand, the second approach specifically blocks ε-amino groups on K residues using *O*-methylisourea, leaving the α-amino group of a protein to be tagged with biotin (Yoshihara *et al.*, 2008).

Contrary to the above two approaches, Subtiligase does not require any modification of K ε-amino groups. This approach utilises an engineered enzyme (i.e. subtiligase) to directly transfer a biotin-tagged peptide ester to the N-terminus of a protein in a single step (Yoshihara *et al.*, 2008). After proteolysis, all three approaches employ immobilised avidin or streptavidin beads to capture the biotin-tagged peptides. The biotin tags used in these

approaches all contain a cleavable linker that allows subsequent release of the captured peptides for liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis.

In theory, the vast majority of protein N-termini (~ 80 %) are discarded by these positive selection strategies due to endogenous N-terminal (Nt)-acetylation. The blocked N-terminal peptides will not be tagged with biotin nor captured by avidin/streptavidin beads during AP. As a result, the biotin-tagged protein N-termini will in principle be detected by LC-MS/MS with a higher opportunity. Furthermore, positive selection strategies are not hampered by the limited binding capacity of coupling reagents (e.g. NHS-activated Sepharose) since they do not require the binding and removal of abundant internal peptides. Although such strategies have advantages over their negative selection counterparts in terms of detection sensitivity and false-positive rates, their feasibility is still limited by multiple factors including sample size and reagent availability. For instance, the Subtiligase approach requires a large quantity of starting materials (i.e. ~ 100 mg of proteins) due to the relatively low tagging efficiency of this enzyme. Additionally, the proprietary enzyme and its tag substrate are not locally available. The advantages and disadvantages of each positive selection strategy are described in Table 4.1.

Table 4.1 Comparison of three major positive selection strategies with respect to their strength and weakness[a].

N-terminalomics	Strength	Weakness
N-CLAP[b]	Widely available and inexpensive reagents	Extensive chemical modifications may lead to significant sample loss, indirect inference of true protein N-termini, off-target N-terminal tagging may lead to false-positives
***O*-methylisourea**	Relatively inexpensive reagents	Complete lysine blocking may be difficult to achieve, off-target N-terminal tagging may lead to false-positives
Subtiligase	Absolute specificity towards α-amines at the N-termini, single-step tagging of free N-termini that minimises sample loss, short handling time	Relies on a proprietary enzyme with limited availability, requires a large quantity of proteins, limited reaction scope leads to false-negatives

[a] In general, positive selection strategies are more suitable to identify protease substrates instead of protein N-termini. All three strategies rely on the match of a single peptide (the N-terminal peptide) for protein identification. [b] N-CLAP: N-terminalomics by chemical labeling of the α-amine of proteins.

Given the current states of these approaches, we explored the feasibility of an alternative positive selection strategy based on another chemical reaction that is specific for α-amino groups. Selective transamination, mainly developed by H. B. Dixon in the 1960s, is the selective conversion of α-amino groups at protein N-termini to carbonyl groups (C=O) using a combination of small molecules (Dixon and Fields, 1972). In contrast, ε-amino groups are not modified as this reaction requires the presence of a peptide bond adjacent to the target amine. After selective transamination, the modified N-terminal residues were removed using a carbonyl-reactive compound to study the function of protein N-termini (Dixon and Moret, 1964, Dixon and Moret, 1965). Although not pursued by Dixon, the carbonyl group introduced by this reaction is highly reactive and can be exploited for affinity tagging (e.g. by biotinylation). Recently, selective transamination has been utilised to block free protein N-termini of model proteins with carbonyl-reactive compounds. After proteolysis with LysN, the blocked N-terminal peptides were negatively selected by amine-reactive glass (Sonomura *et al.*, 2009a). To date, however, this reaction has not been applied to affinity tag free protein N-termini for positive selection.

The present study aimed to develop a novel positive selection approach by combining selective transamination with carbonyl-specific tagging and AP. Similar to the critical evaluation and optimisation of the NHS-Sepharose approach (see Chapter 3), this positive selection approach was also established and tested on samples with an increasing complexity: from synthetic peptides to model proteins.

4.2 Results

4.2.1 Design of a transamination-based positive selection strategy

In addition to the reaction developed by Dixon (1964), several transamination routes have been proposed more recently by M. B. Francis (2006, 2010). These reactions use either pyridoxal-5-phosphate (PLP) or *N*-methylpyridinium-4-carboxaldehyd (RS) to achieve the selective conversion of α-amino groups to carbonyl groups. However, the reaction developed by Dixon was chosen for the following reasons: I. the rate of transamination is higher using this method than the PLP-mediated reaction; II. this method can be applied to modify a wide variety of N-terminal residues whereas the RS-mediated reaction is largely limited to proteins with an N-terminal glutamate (E) residue; III. the reagents required by this method are readily available and more affordable.

As illustrated in Figure 4.1, selective transamination of protein N-terminal residues involves three major components: an amino acceptor, a heavy metal ion as the catalyst, and a high concentration of base. The amino acceptor of choice is usually glyoxylate, whereas copper(II) sulfate is the preferred catalyst for this reaction. Initially sodium acetate was selected as the base, but it could be substituted by pyridine, especially for peptide transamination (Dixon and Fields, 1972). The reaction typically proceeds at pH 6 for 0.5 – 2 hours (h) at room temperature (Dixon and Fields, 1972, Papanikos *et al.*, 2001, Sonomura *et al.*, 2009a).

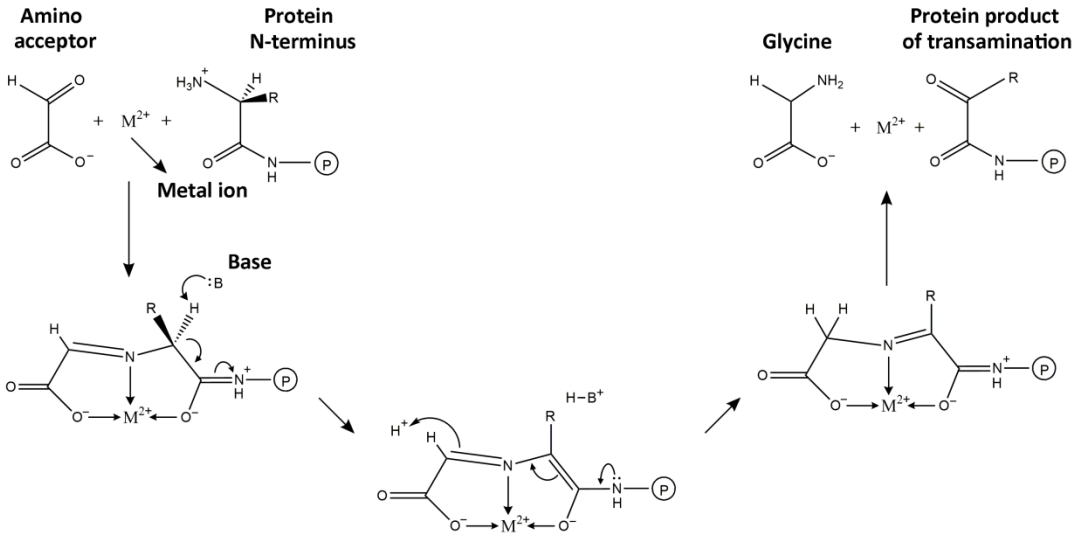


Figure 4.1 Schematic illustration of selective transamination of protein or peptide N-termini (modified from Dixon, 1984). The negatively charged oxygen of an amino acceptor (e.g. glyoxylate) is first juxtaposed with the carbonyl oxygen of the first amino acid residue of a protein or peptide Ⓟ via a divalent metal cation. The presence of a base (B) allows the formation of an imine which, following a sequence of proton transfer events, is hydrolysed thereby converting the α-amine to a reactive carbonyl group.

In the present experiments, peptides or proteins needed to be separated from other reactants after selective transamination for further modifications. Due to the differences in their physico-chemical properties, the modified proteins could be purified by gel-filtration chromatography whereas the modified peptides were only compatible with reversed-phase chromatography. The purified peptides/proteins were then subjected to affinity tagging of the newly introduced carbonyl groups. Biotin was selected as the affinity tag in the current strategy owing to the strength and selectivity of biotin-avidin interaction (Dundas *et al.*, 2013). In addition, a biotin tag can be tailored to specific tagging of the transaminated protein N-termini when linked to a carbonyl-reactive group. Two types of carbonyl-specific biotin were tested: hydrazide-based and alkoxyamine-based. Despite the differences in their functional groups, both types are commercially available for tagging carbonyl groups. In addition to the biotin tags, *O*-benzylhydroxylamine ($BnONH_2$) was also employed in the current study since it is routinely used to block carbonyl groups that are introduced by different transamination routes (Gilmore *et al.*, 2006, Witus *et al.*, 2013).

After biotin tagging or the modification with $BnONH_2$, the peptide samples were conveniently analysed by LC-MS/MS to identify such modifications. In contrast, an immunoblotting approach (e.g. dot blotting) greatly simplified the detection of the biotin tag on intact proteins by use of an antibody against the biotin moiety. Next, the biotin-tagged proteins were either directly subjected to AP or digested with a protease to yield a mixture of peptides. Two variants of immobilised avidin beads were selected for the AP experiments: NeutrAvidin agarose and monomeric avidin agarose. As described in Chapter 1, the former binds to biotin with minimised nonspecific interactions (Marttila *et al.*, 2000), whereas the latter only requires mild conditions for elution (Henrikson *et al.*, 1979).

For protein AP, the biotin-tagged proteins were directly treated with the immobilised avidin beads without protease digestion. The outcome of protein AP was conveniently evaluated using sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). For AP at the peptide level, different proteases were chosen depending on the nature of the biotin-tagged proteins: Glu-C for chicken egg-white lysozyme (lysozyme C) and trypsin for bovine serum albumin (BSA). After protease digestion, the resulting peptides could be directly analysed by LC-MS/MS to locate the biotin tag. Alternatively, the peptide samples were incubated with either avidin resin to enrich for the biotin-tagged peptides. Finally, the enriched peptides were subjected to LC-MS/MS analysis for evaluating the outcome of peptide AP. The entire workflow of the selective transamination approach is summarised in Figure 4.2.

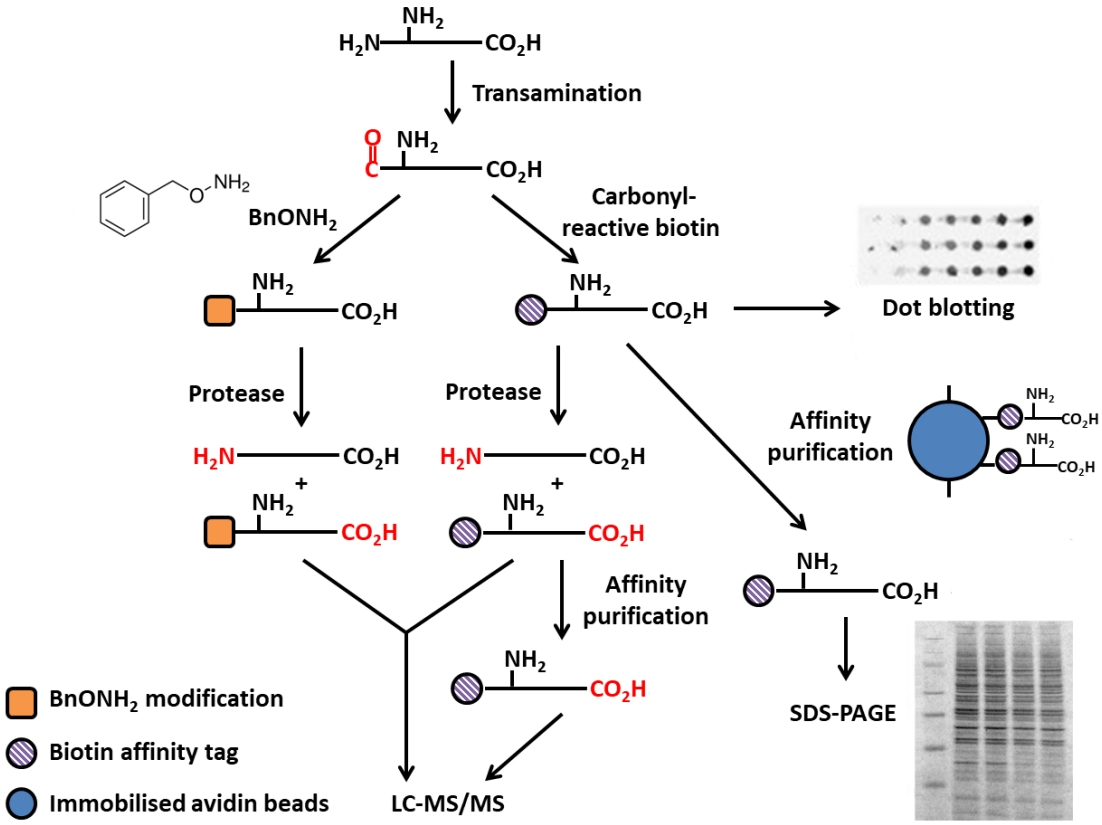


Figure 4.2 Schematic of a positive selection strategy for N-terminalomic studies (modified from Wehr and Levine, 2012). The proposed strategy is based on selective transamination and consists of five steps: I. selective transamination of protein N-termini; II. biotinylation or modification with *O*-benzylhydroxylamine ($BnONH_2$); III. protease digestion; IV. affinity purification (AP) using immobilised avidin resin; V. liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis. Dot blotting and sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) were employed to simplify the evaluation of carbonyl-specific biotinylation and AP, respectively. Peptide transamination was directly analysed by LC-MS/MS without protease digestion. The red labels indicate new chemical groups introduced by selective transamination or protease digestion.

4.2.2 Experiments with model peptides

Selective transamination was first tested on synthetic peptides, which do not require protease digestion in order to be detected by LC-MS/MS. There were two synthetic peptides available at the time when these experiments were initiated: human adrenocorticotropic hormone (ACTH) and rat renin substrate. Human ACTH is a decapeptide (amino acid sequence: SYSMEHFRWG), whereas rat renin substrate is composed of 14 amino acid residues (DRVYIHPFHLLYYS). These two peptides were subjected to selective transamination that in theory introduces a reactive carbonyl group. The transaminated peptides were then treated with an alkoxyamine-based biotin tag, which is specific for carbonyl groups and hence should be attached to the N-termini of the transaminated peptides. In parallel to the biotinylation experiments, the carbonyl-reactive compound $BnONH_2$ was also employed to modify the transaminated peptides. The purpose of this treatment was to independently evaluate peptide transamination and validate the results of biotinylation.

Through LC-MS/MS analyses, these two peptides were examined in detail to assess the extent of the transamination reaction, the position of the newly introduced carbonyl group, and the chemical reactivity of this group. Such analyses were conducted on peptide samples at various time points: before transamination, after transamination, and after the treatment with $BnONH_2$/biotin. As described in Chapter 2, peak lists were compiled from Orbitrap RAW data and then searched against the decoy sequence of each peptide (SYSMEHFRWG or DRVYIHPFHLLYYS). Variable modifications including Nt-transamination (-1.03 Da), N-terminal tagging with $BnONH_2$ (+104.03 Da), or Nt-biotinylation (+825.33 Da) were set in order to identify these N-terminal modifications.

Prior to selective transamination, both human ACTH and rat renin substrate were directly analysed by LC-MS/MS for quality control. Human ACTH was eluted at 35.67 minute (min) during LC separation (1-h gradient), whereas rat renin substrate was eluted at 38.74 min (retention time, *RT*). Both peptides were detected by MS as a single peak (Figure 4.3).

The corresponding full mass spectra were inspected for peptide identification. For human ACTH, a full mass spectrum (*RT* = 35.83 min) exhibited two major peaks: the peak with the higher intensity was a doubly charged ion with a mass-to-charge ratio (*m*/*z*) of 650.28, whereas the peak with the lower intensity was a triply charged ion at *m*/*z* = 433.86 (Figure 4.4A). By calculation, both peaks were derived from the same molecule, with a monoisotopic mass of 1298.56 Da. It was equal to the expected molecular weight (MW) of the peptide human ACTH.

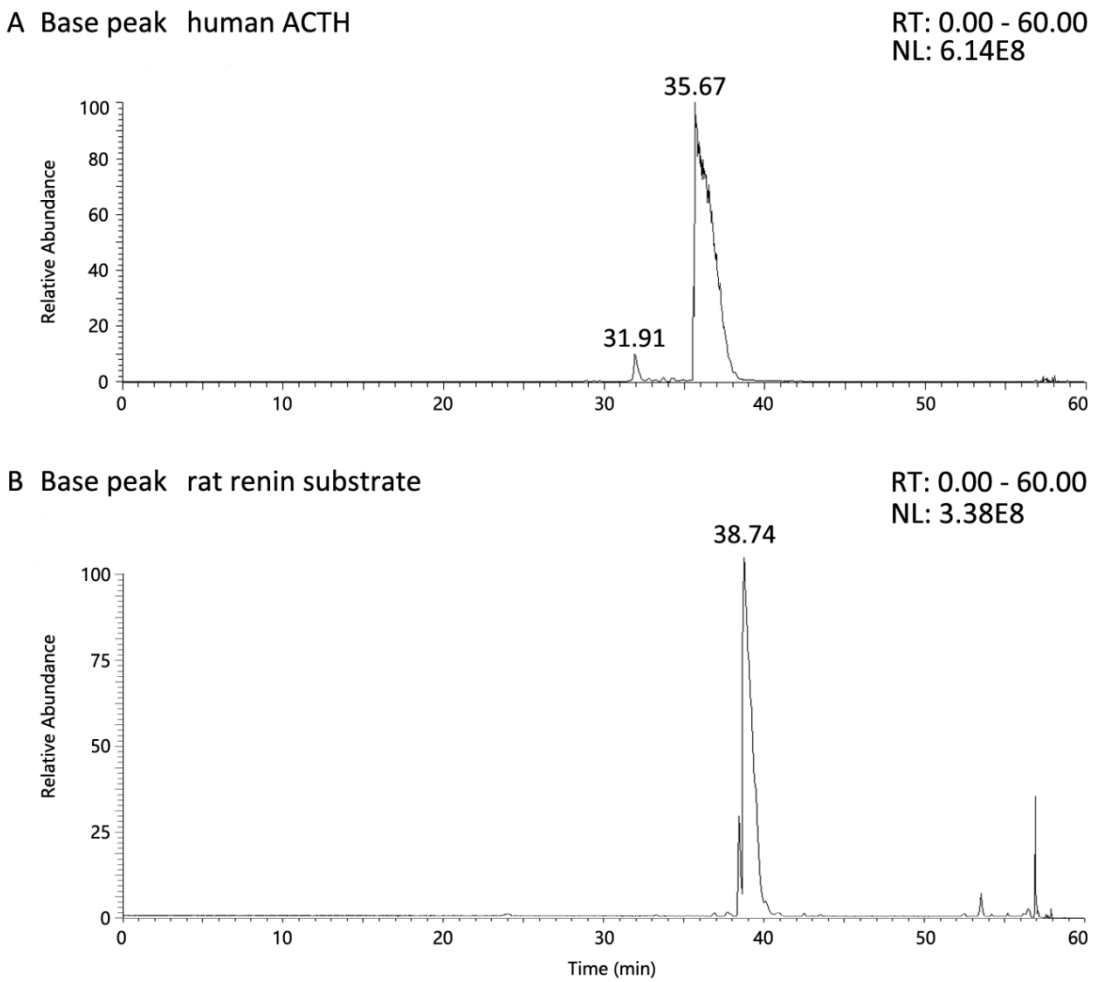


Figure 4.3 Base peak chromatograms of the model peptides human ACTH (**A**) and rat renin substrate (**B**). Both peptides were detected as a single peak in LC-MS/MS analyses. ACTH: adrenocorticotropic hormone; LC-MS/MS: liquid chromatography–tandem mass spectrometry; *RT*: retention time; NL: normalised intensity level; min: minute.

Similarly, the full mass spectrum (*RT* = 38.74 min) of rat renin substrate also revealed two major peaks with *m*/*z* of 608.31 and 911.97, respectively (Figure 4.4B). Both ions were derived from the same molecule with a monoisotopic mass of 1821.92 Da, which matched the expected MW of rat renin substrate.

In addition to manual inspection of the full mass spectra, MS/MS data of human ACTH and rat renin substrate were processed and searched against their respective peptide sequences using the Mascot software. As shown in Table 4.2, both peptides were successfully identified by Mascot database searches. Therefore, the two peptide samples were confirmed as human ACTH (SYSMEHFRWG) and rat renin substrate (DRVYIHPFHLLYYS).

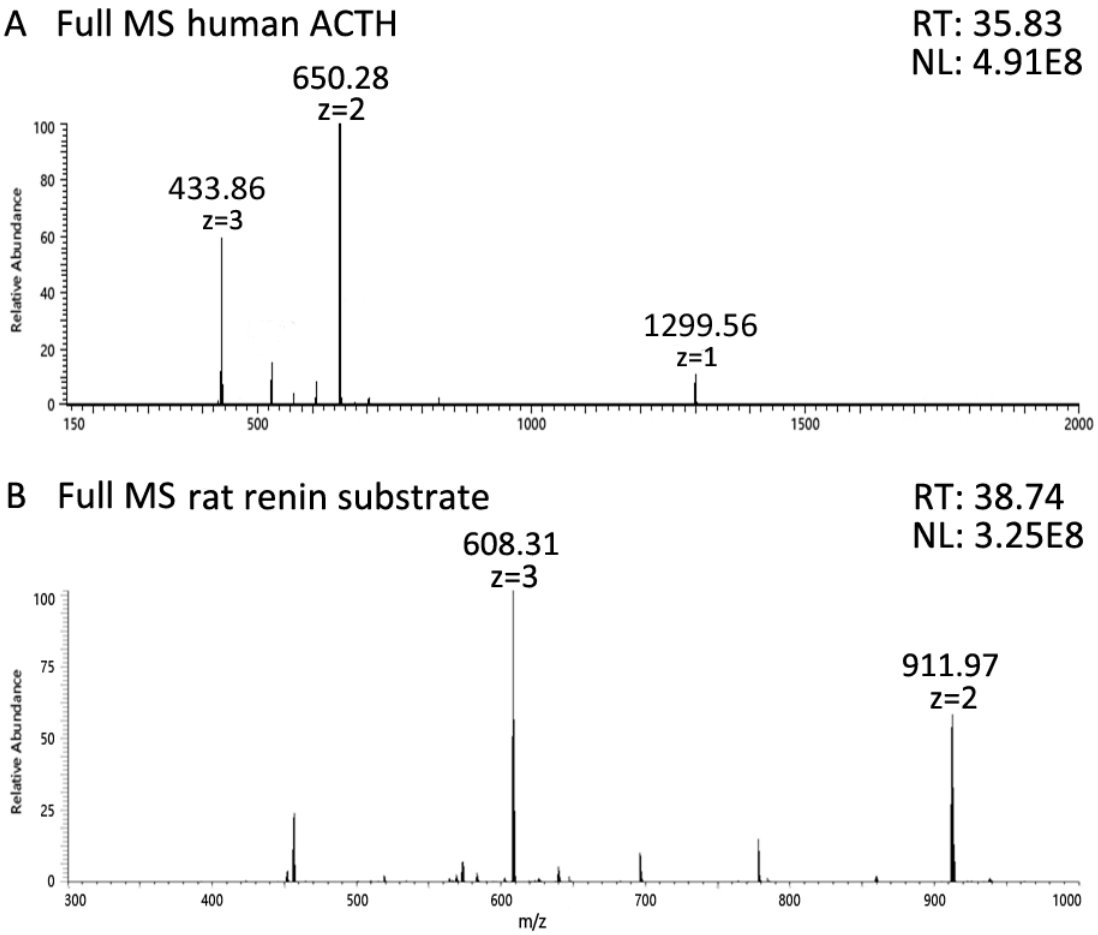


Figure 4.4 Full mass spectra of human ACTH (*RT* = 38.74 min) and rat renin substrate (*RT* = 35.93 min). For human ACTH (**A**), two major peaks are the doubly charged ion at *m/z* = 650.28 and the triply charged ion at *m/z* = 433.86. For rat renin substrate (**B**), two major peaks are the doubly charged ion at *m/z* = 911.97 and the triply charged ion at *m/z* = 608.31. ACTH: adrenocorticotropic hormone; *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio; min: minute.

Table 4.2 Mascot search results of the peptides human ACTH and rat renin substrate[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value	Duplicate PSM No.
Human ACTH					
SYSMEHFRWG (1 – 10)	650.2816	1298.5502	60	9.9E-7	19
SYSMEHFRWG (1 – 10) + Oxidation (M)	658.2791	1314.5452	50	1.0E-5	23
Rat renin substrate					
DRVYIHPFHLLYYS (1 – 14)	608.3150	1821.9202	31	0.00078	23

[a] A single PSM is shown to represent each significantly identified peptide (*E*-value ≤ 0.05). ACTH: adrenocorticotropic hormone; *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

After the quality control, these two peptide samples were subjected to selective transamination. This reaction was carried out under two different conditions ("salt-free" or "salt-based", see section 2.8), and both yielded the same reaction products as revealed by LC-MS/MS analyses. The "salt-free" condition was selected for all subsequent transamination experiments due to its compatibility with both peptide and protein samples. Therefore, only the data obtained using the "salt-free" transamination are described below. Base peak chromatograms showed that the transamination of human ACTH or rat renin substrate resulted in multiple peaks during LC separation (1-h gradient). For the reaction products of human ACTH, three major peaks were present at *RT* = 41.11, 41.85, and 42.99 min (Figure 4.5A). In contrast, selective transamination of rat renin substrate yielded two major peaks at *RT* = 42.21 and 44.35 min, respectively (Figure 4.5B).

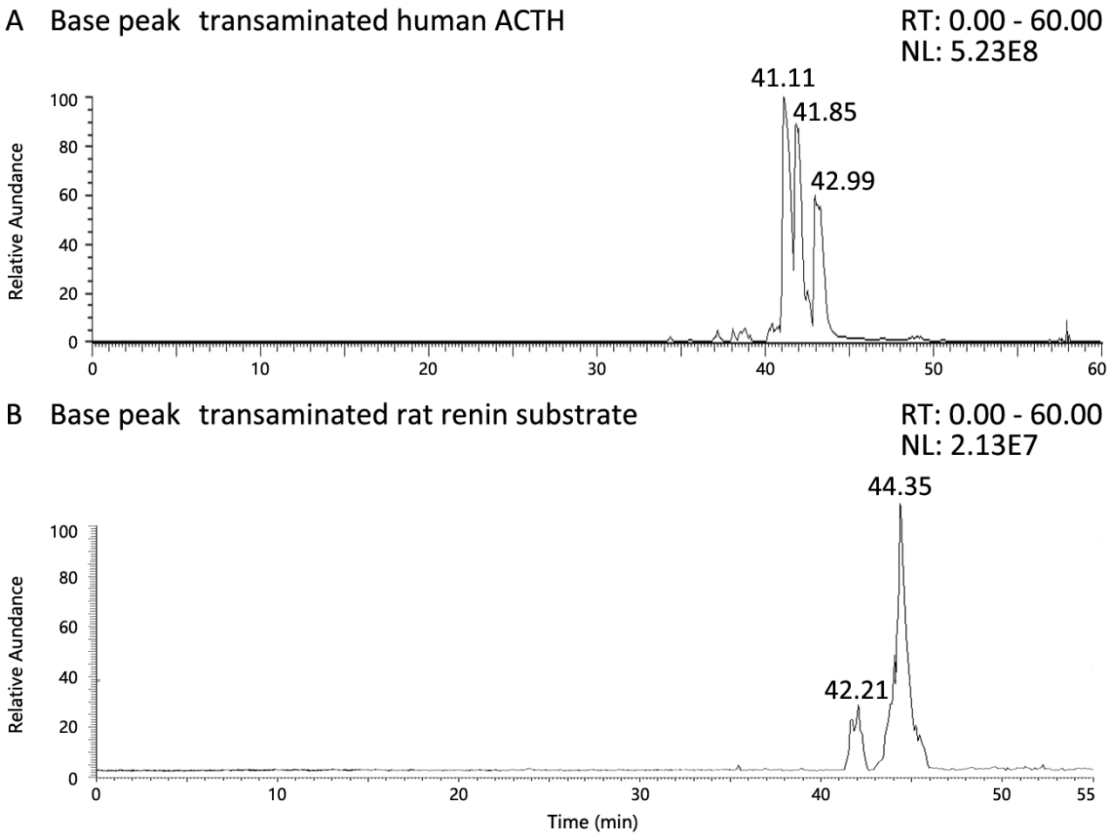


Figure 4.5 Base peak chromatograms of human ACTH (**A**) and rat renin substrate (**B**) after transamination. Peaks in the chromatograms may correspond to the reaction products of each peptide. ACTH: adrenocorticotropic hormone; *RT*: retention time; NL: normalised intensity level; min: minute.

These peaks were further analysed using their corresponding full mass spectra. For human ACTH, the expected transamination product (MW: 1297.52 Da) was detected as a doubly charged ion at *m/z* = 649.77 (Figure 4.6A). In contrast, a high-intensity ion at *m/z* = 687.28 (*z* = 2) was also present. By calculation, the high-intensity ion was derived from a by-product with a MW of 1372.54 Da, corresponding to a mass shift of +74 Da. The intensities of these two ions were then compared using extracted ion chromatograms (XIC), which only monitor the specified targets. As shown in Figure 4.6B, the correct transamination product was only 20 % of the +74 Da by-product in terms of signal intensity.

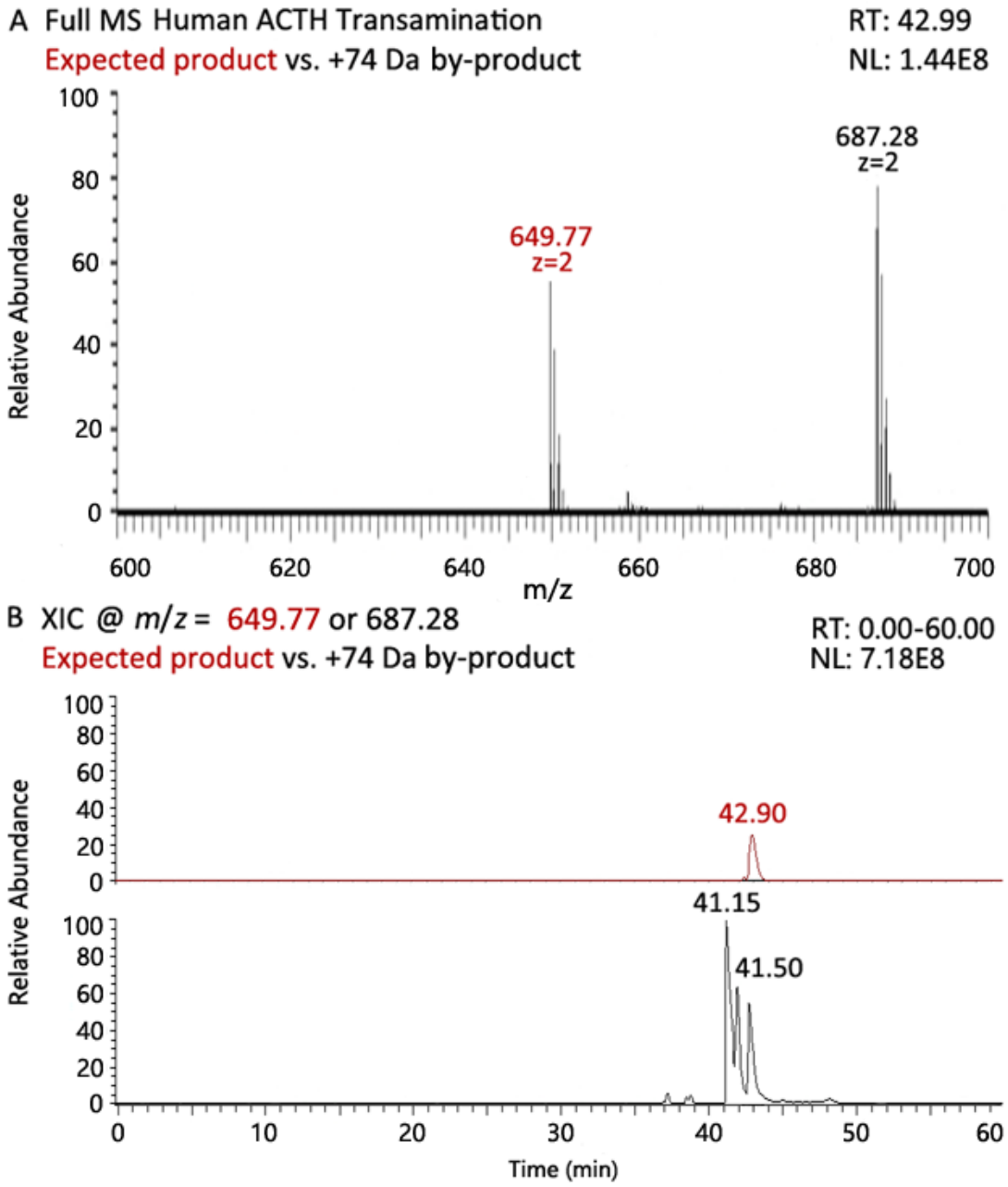


Figure 4.6 Full mass spectrum (**A**) and extracted ion chromatogram (XIC; **B**) of the expected transamination product (red) and the +74 Da by-product (black) of human ACTH (adrenocorticotropic hormone). *m/z:* mass-to-charge ratio; *RT*: retention time; NL: normalised intensity level; min: minute.

Following manual inspection of the full mass spectra, the MS/MS data of the reaction products were subjected to a Mascot search against the sequence of human ACTH. As shown in Table 4.3, the correct reaction product was successfully identified by the Mascot search. However, native human ACTH was also identified, which suggested that selective transamination did not proceed to completion. In contrast, the +74 Da by-product was not identified since this unknown modification was not included in the search parameters.

Table 4.3 Result of the Mascot database search for the transamination products of human ACTH[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value	Duplicate PSM No.
SYSMEHFRWG (1 – 10)	650.2859	1298.5502	64	3.7E-7	7
SYSMEHFRWG (1 – 10) + Oxidation (M)	658.2817	1314.5452	44	4.4E-5	5
SYSMEHFRWG (1 – 10) + Transamination (N-term)	649.7687	1297.5186	61	8.8E-7	11

[a] A single PSM is shown to represent each significantly identified peptide (*E*-value ≤ 0.05). ACTH: adrenocorticotropic hormone; *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

For rat renin substrate, three major reaction products were detected in the full mass spectra. As shown in Figure 4.7, the three peaks were detected as triply charged ions at *m/z* = 593.31, 632.98, and 607.98 (in the order of decreasing signal intensity). By calculation, the peak with the highest intensity (*m/z* = 593.31) was derived from a side product with a mass shift of -45 Da. This product was likely produced by decarboxylation of the N-terminal aspartate (D) residue of rat renin substrate, which will be discussed in section 4.3. Consistent with the transamination of human ACTH, a +74 Da by-product (*m/z* = 632.98) was also detected with comparatively lower intensity. Finally, the peak with the lowest intensity (*m/z* = 607.98) corresponded to the expected reaction product (MW: 1820.77 Da). The correct reaction product was only one seventh of the decarboxylated peptide in terms of signal intensity.

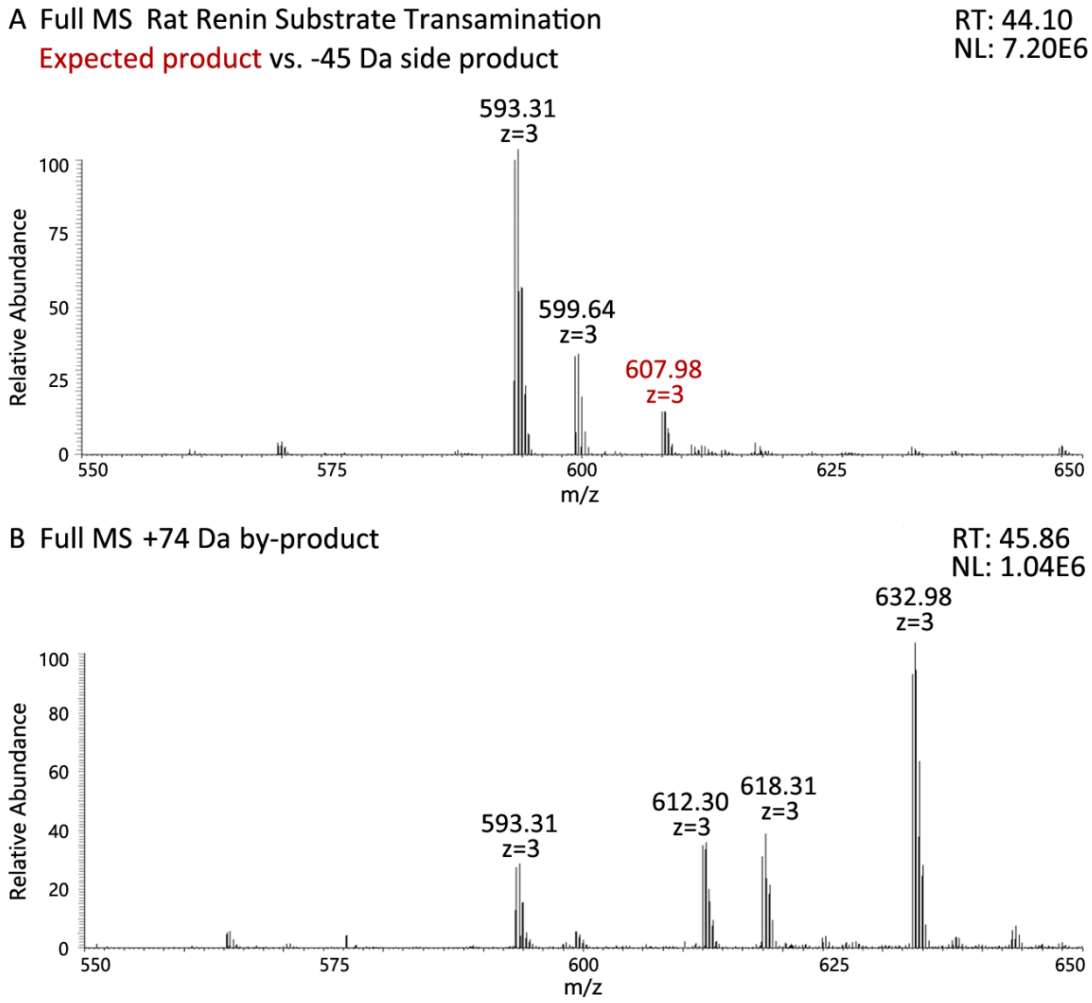


Figure 4.7 Full mass spectra of the expected transamination product and major by-products of rat renin substrate. (**A**) Full mass spectrum (*RT* = 44.10 min) of the correct product (*m/z* = 607.98, *z* = 3) and the -45 Da side product (*m/z* = 593.31, *z* = 3). (**B**) Full mass spectrum (*RT* = 45.86 min) of the +74 Da by-product (*m/z* = 632.98, *z* = 3). *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio; min: minute.

The next step in the current experiments was to further modify the transaminated human ACTH and rat renin substrate with $BnONH_2$. The aim of this treatment was to ascertain the nature and reactivity of the transamination products. For human ACTH, the treatment with this compound resulted in the presence of two peaks at *RT* = 49.56 and 51.13 min, respectively. These peaks are shown in an XIC (Figure 4.8A). The corresponding full mass spectra at both time points showed the same doubly charged ion at *m/z* = 702.30 (Figure 4.8B). This ion was determined to result from the modified peptide with a mass shift of +104.03 Da. It matched the expected value of $BnONH_2$ modification following selective transamination.

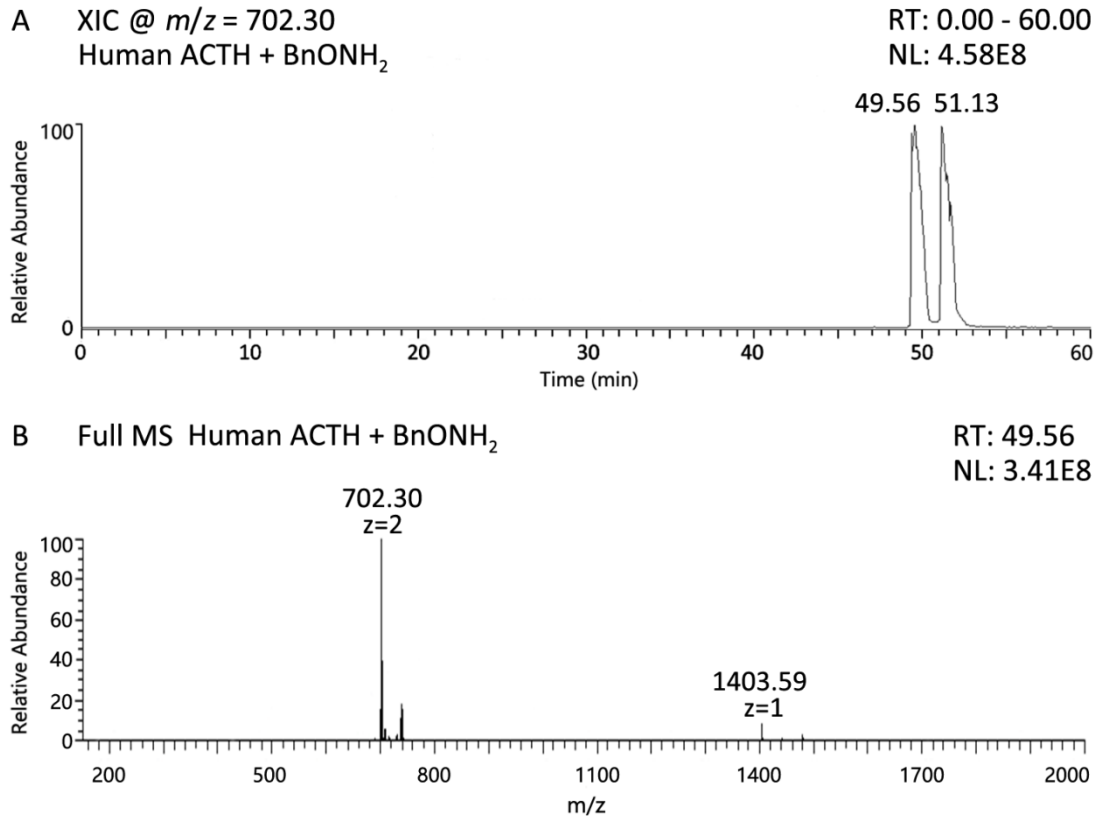


Figure 4.8 (**A**) Extracted ion chromatogram (XIC) of human ACTH that was modified by successive transamination and oxime formation with $BnONH_2$. (**B**) Corresponding full mass spectrum of the $BnONH_2$-modified peptide (*m/z* = 702.30, *z* = 2) at *RT* = 49.56 min. ACTH: adrenocorticotropic hormone; $BnONH_2$: *O*-benzylhydroxylamine; *m/z*: mass-to-charge ratio; *RT*: retention time; NL: normalised intensity level; min: minute.

For rat renin substrate, the treatment with this compound yielded two peaks at *RT* = 51.44 and 52.88 min, respectively (Figure 4.9). Both peaks were shown by the corresponding full mass spectra to be a triply charged ion at *m*/*z* = 642.99 (Figure 4.10). This ion corresponded to a MW of 1925.95 Da, which was the expected value of the $BnONH_2$-modified correct transamination product of rat renin substrate. This peptide was accompanied by an additional pair of peaks (*RT* = 51.44 and 52.40 min, respectively), which were detected as a triply charged ion at *m*/*z* = 628.33. This ion might correspond to the decarboxylated peptide that was further modified with $BnONH_2$.

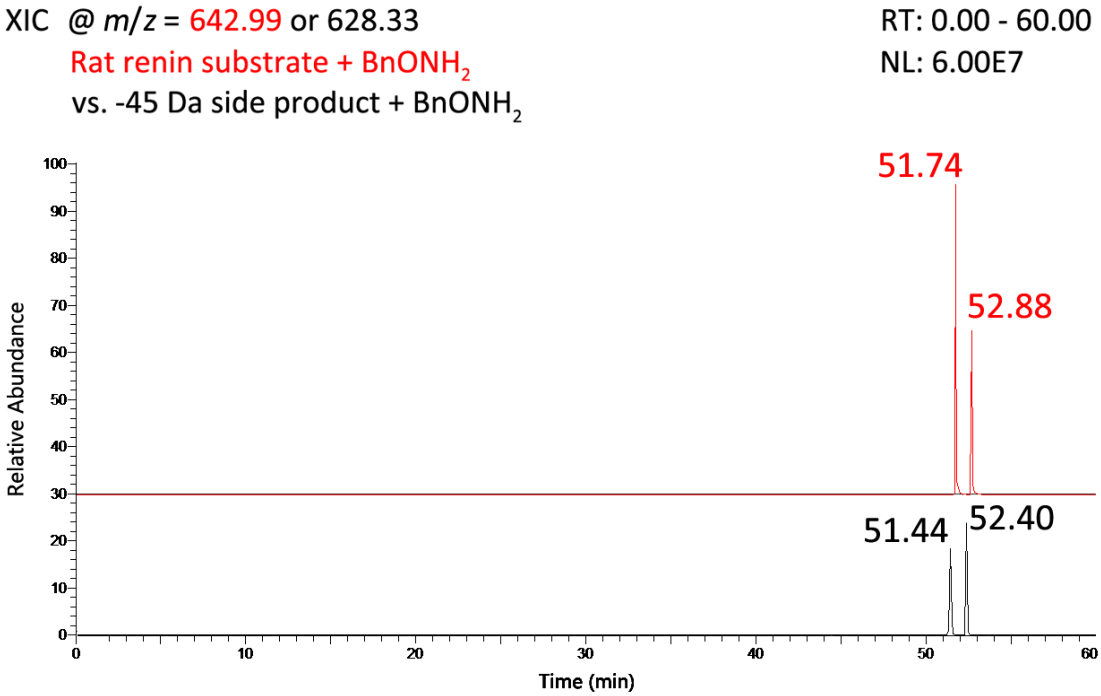


Figure 4.9 Combined extracted ion chromatograms (XIC) of rat renin substrate, which was modified by successive transamination and oxime formation with $BnONH_2$. The $BnONH_2$-modified correct transamination product (*m*/*z* = 642.99, *z* = 3) is shown in read, whereas the $BnONH_2$-modified -45 Da side product (*m*/*z* = 628.33, *z* = 3) is shown in black. $BnONH_2$: *O*-benzylhydroxylamine; *m*/*z*: mass-to-charge ratio; *RT*: retention time; NL: normalised intensity level; min: minute.

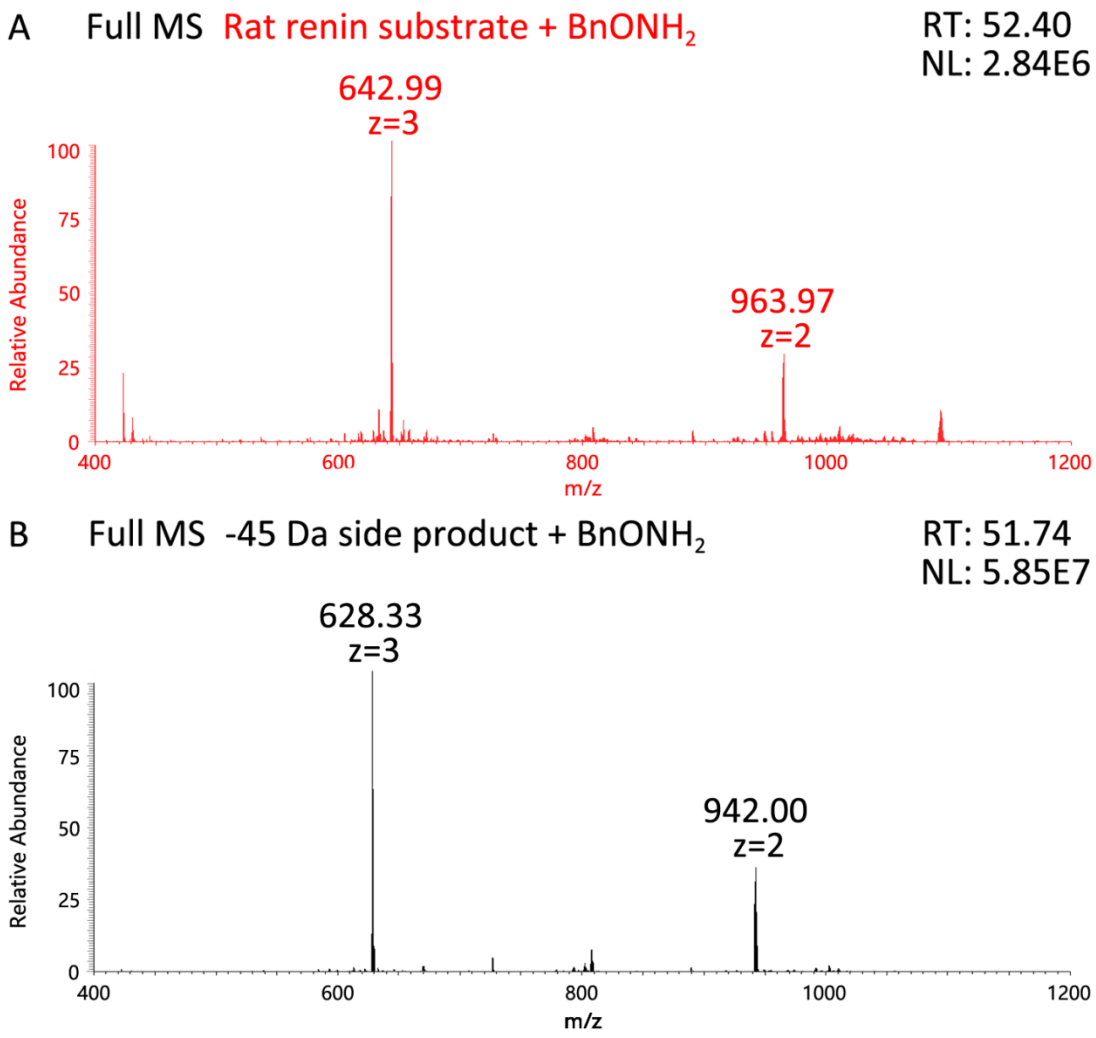


Figure 4.10 Full mass spectra of rat renin substrate that was modified by successive transamination and oxime formation with $BnONH_2$. (**A**) The $BnONH_2$-modified correct transamination product (*m/z* = 642.99 or 963.97; in red). (**B**) The $BnONH_2$-modified decarboxylated peptide (*m/z* = 628.33 or 942.00; in black). $BnONH_2$: *O*-benzylhydroxylamine; *m/z*: mass-to-charge ratio; *RT*: retention time; NL: normalised intensity level.

The MS/MS data of the $BnONH_2$-modified transamination products were then subjected to Mascot searches against the corresponding peptide sequence. This established that the $BnONH_2$ modification was located at the N-terminus of both the transaminated human ACTH and rat renin substrate (Table 4.4). Taken together, these results showed that the correct transamination products of human ACTH and rat renin substrate were both amenable to the carbonyl-specific modification with $BnONH_2$. Therefore, both correct transamination products possessed a reactive carbonyl group at their N-termini. Furthermore, certain side products of transamination (e.g. decarboxylated rat renin substrate) might also exhibit the reactivity attributed to a carbonyl group.

Table 4.4 Results of the Mascot database searches for $BnONH_2$-modified human ACTH and rat renin substrate[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value	Duplicate PSM No.
$BnONH_2$-modified human ACTH					
SYSMEHFRWG (1 – 10)	**650.2870**	**1298.5502**	**64**	**3.8E-7**	**8**
SYSMEHFRWG (1 – 10) + $BnONH_2$ (N-term)	**702.2975**	**1402.5764**	**56**	**2.7E-6**	**7**
$BnONH_2$-modified rat renin substrate					
DRVYIHPFHLLYYS (1 – 14)	**608.3141**	**1821.9202**	**24**	**0.0042**	
DRVYIHPFHLLYYS (1 – 14) + $BnONH_2$ (N-term)	**642.9850**	**1925.9464**	**17**	**0.02**	**2**

[a] A single PSM is shown for each significantly identified peptide (*E*-value ≤ 0.05). ACTH: adrenocorticotropic hormone; $BnONH_2$: *O*-benzylhydroxylamine; *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

Having confirmed that the expected transamination product of both human ACTH and rat renin substrate possessed a reactive carbonyl group, they were then subjected to affinity tagging with alkoxyamine-based biotin (i.e. alkoxyamine-PEG_4-SS-PEG_4-biotin). Consistent with the $BnONH_2$ modification, two adjacent peaks (*RT* = 49.47 and 50.38 min, respectively) were present in the XIC of the biotinylated human ACTH (Figure 4.11A). These two peaks contained a triply charged ion (*m/z* = 708.97) and a doubly charged ion (*m/z* = 1062.94) that both corresponded to a reaction product with a MW of 2123.89 Da (Figure 4.11B). In comparison with the native human ACTH, this product showed the expected mass shift (+825.33 Da) due to successive transamination and biotin tagging.

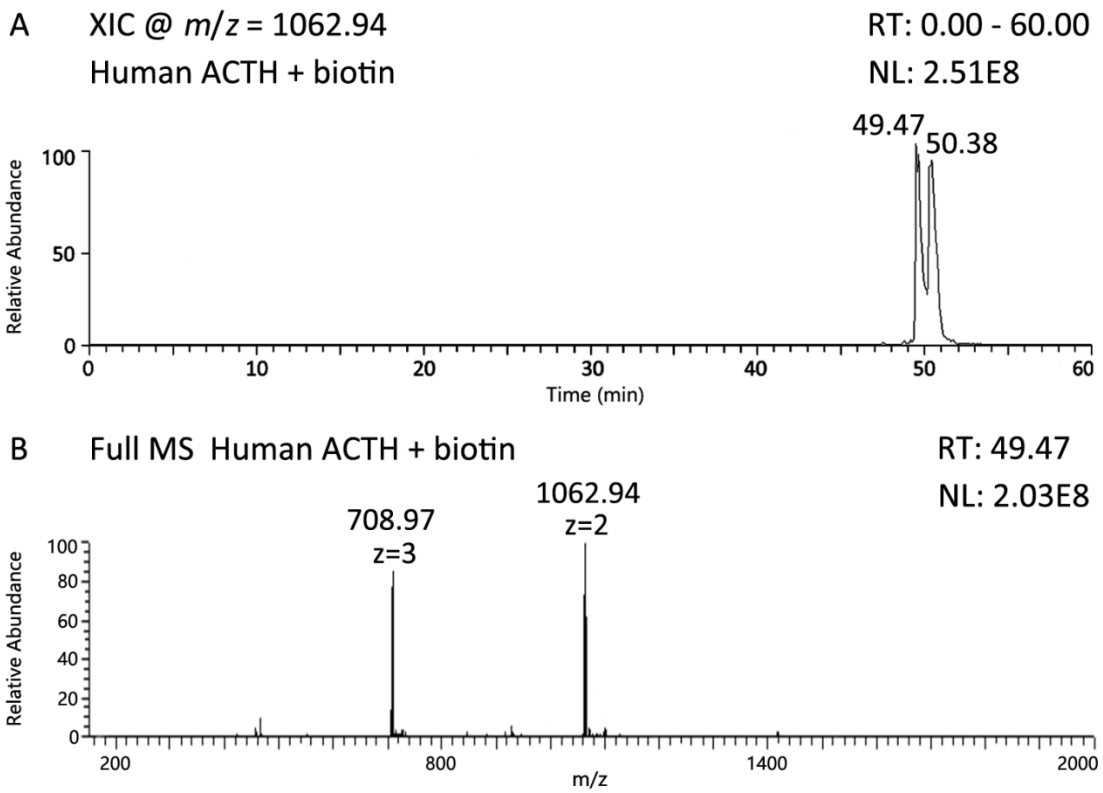


Figure 4.11 Extracted ion chromatogram (XIC; **A**) and full mass spectrum (**B**) of the biotinylated human ACTH (adrenocorticotropic hormone). *RT*: retention time; NL: normalised intensity level; *m*/*z*: mass-to-charge ratio.

The position of the attached biotin tag was also confirmed by a Mascot database search, with Nt-biotinylation (+825.33 Da) included in the variable modifications (Table 4.5). After Higher-energy Collisional Dissociation (HCD) fragmentation of the precursor ion of the biotinylated human ACTH (*m*/*z* = 1062.94, *z* = 2), *y*-series fragment ions (y_3 – y_8) were identified in the tandem mass spectrum (Figure 4.12 & Table 4.6). These *y*-ions were identical to those derived from the native human ACTH, strongly suggesting that the biotin tag had been attached to the N-terminus of the peptide. In addition, the identified *y*-ions were accompanied by an intense peak at *m*/*z* = 270.13 (*z* = 1). This peak likely corresponded to a fragment ion that resulted from putative cleavage of the attached biotin tag itself during HCD fragmentation (discussed in section 4.3).

Nonetheless, the Mascot search result indicated that neither transamination nor biotin tagging proceeded to completion, since both the native and transaminated human ACTH were also detected in the biotin-tagged sample. Meanwhile, the biotinylated rat renin substrate could not be directly identified by a Mascot search using the same parameters. An error-tolerant search was thus performed to account for possible side reactions in

transamination, which would confer additional mass shifts to the peptide. The logic behind the error-tolerant search was that the decarboxylated rat renin substrate was detected with high intensity and this side product likely possessed a reactive carbonyl group as well.

Table 4.5 Result of the Mascot database search for biotinylated human ACTH[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	Score	*E*-value	Duplicate PSM No.
SYSMEHFRWG (1 – 10)	650.2870	1298.5502	60	9.4E-7	4
SYSMEHFRWG (1 – 10) + Biotin (N-term)	1062.9454	2123.8825	21	0.0084	2
SYSMEHFRWG (1 – 10) + Transamination (N-term)	649.7655	1297.5186	52	6.5E-6	5

[a] A single PSM is shown for each significant peptide (*E*-value ≤ 0.05). ACTH: adrenocorticotropic hormone; *m/z*: mass-to-charge ratio; MW: molecular weight; *E*-value: peptide expectation value; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

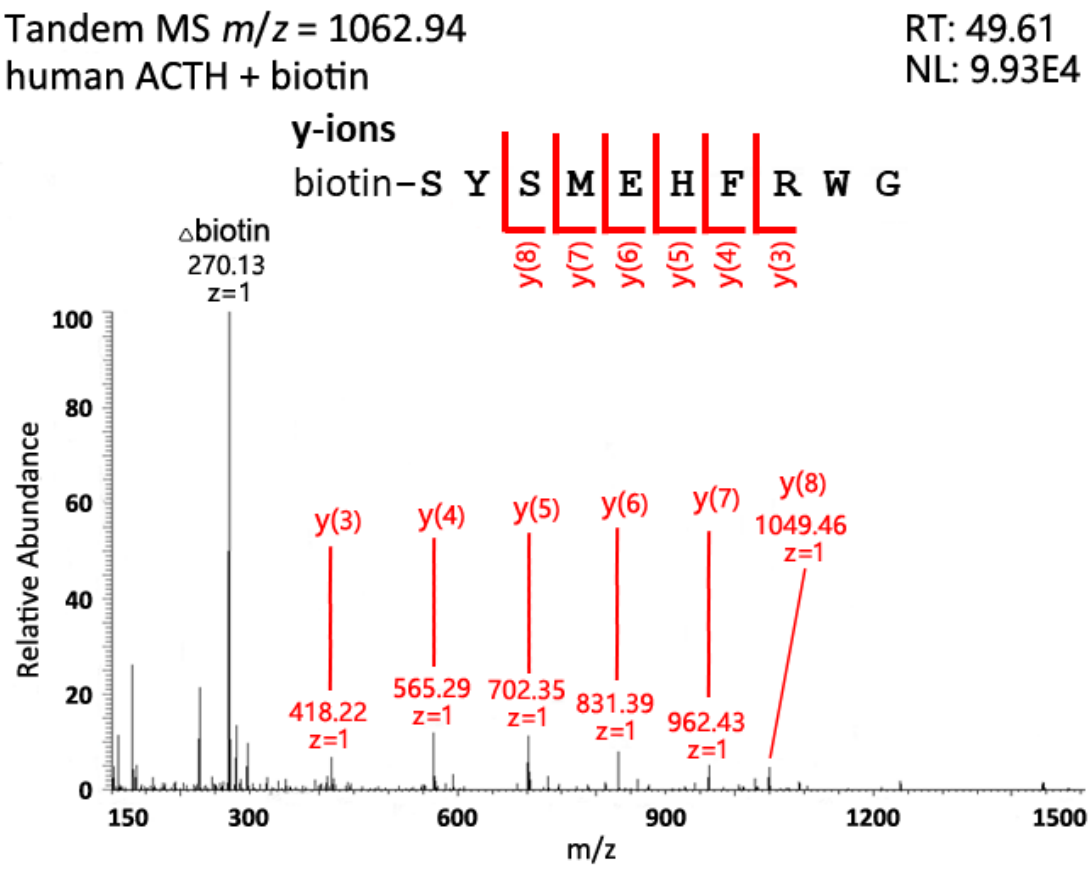


Figure 4.12 Tandem mass spectrum of biotinylated human adrenocorticotropic hormone (ACTH, amino acid sequence: SYSMEHFRWG). The expected *m/z* values for all possible fragment ions are shown in Table 4.6. The annotated fragment ions are labelled in red. The fragment ion at *m/z* = 270.13 (*z* = 1), which likely resulted from putative cleavage of the attached biotin tag, is denoted by $_{\Delta}$biotin. *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio.

Table 4.6 Fragment ions (*in silico* predicted) of the biotin-tagged human ACTH (amino acid sequence: SYSMEHFRWG). Fragment ions matched to the experimental data are shown in red. ACTH: adrenocorticotropic hormone.

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	913.3715	457.1894			895.3610	448.1841	S							10
2	1076.4349	538.7211			1058.4243	529.7158	Y	1212.5255	606.7664	1195.4989	598.2531	1194.5149	597.7611	9
3	1163.4669	582.2371			1145.4563	573.2318	S	*1049.4622*	525.2347	1032.4356	516.7214	1031.4516	516.2294	8
4	1294.5074	647.7573			1276.4968	638.7520	M	*962.4301*	481.7187	945.4036	473.2054	944.4196	472.7134	7
5	1423.5500	712.2786			1405.5394	703.2733	E	*831.3896*	416.1985	814.3631	407.6852	813.3791	407.1932	6
6	1560.6089	780.8081			1542.5983	771.8028	H	*702.3471*	351.6772	685.3205	343.1639			5
7	1707.6773	854.3423			1689.6667	845.3370	F	*565.2881*	283.1477	548.2616	274.6344			4
8	1863.7784	932.3928	1846.7519	923.8796	1845.7678	923.3876	R	*418.2197*	209.6135	401.1932	201.1002			3
9	2049.8577	1025.4325	2032.8312	1016.9192	2031.8472	1016.4272	W	262.1186	131.5629					2
10							G	76.0393	38.5233					1

The error-tolerant search did identify multiple peptide-spectrum matches (PSMs) that might represent the biotinylated rat renin substrate with additional mass shifts (Table 4.7). Among them, the doubly charged ion at *m*/*z* = 1302.64 was repeatedly detected. Compared to the expected biotinylation product, this putative peptide exhibited a further mass shift of -44 Da. Therefore, potentially it was formed by biotin tagging of the decarboxylated rat renin substrate. Additional search hits included the precursor ions at *m*/*z* = 1295.63 and 1310.64 (*z* = 2 for both). The former corresponded to a further mass shift of -58 Da that likely suggested a complete loss of the side chain (i.e. N-terminal D to G), whereas the latter (-28 Da) likely indicated the conversion from a carboxyl to a hydroxyl group (i.e. N-terminal D to S).

Table 4.7 Result of the Mascot error-tolerant search for biotinylated rat renin substrate[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	ppm	Score	Duplicate PSM No.
Native peptide (Reference)					
DRVYIHPFHLLYYS (1 – 14)	**911.9666**	**1821.9202**	**-0.89**	**54**	**9**
Putative biotinylated peptides					
DRVYIHPFHLLYYS (1 – 14) + Biotin (N-term); -58 Da	**1295.6305**	**2589.2470**	**-0.21**	**27**	
DRVYIHPFHLLYYS (1 – 14) + Biotin (N-term); -44 Da	**1302.6391**	**2603.2627**	**1.07**	**30**	**3**
DRVYIHPFHLLYYS (1 – 14) + Biotin (N-term); -28 Da	**1310.6380**	**2619.2576**	**1.49**	**36**	**1**

[a] A single PSM is shown for each putative biotinylated peptide. The native peptide confidently identified in the same experiment is also included as reference. Peptide score ≥ 13 indicates identity. *m*/*z*: mass-to-charge ratio; MW: molecular weight; ppm: parts per million; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

4.2.3 Experiments with model proteins

Having partially established that peptides could be transaminated and subsequently tagged with biotin, the next step was to perform the same test on two model proteins (lysozyme C and BSA). Previously, these two proteins had already been analysed by LC-MS/MS in their native states (see Chapter 3). Their respective N-terminal peptides were also identified: a 7-residue peptide (KVFGRCE) in lysozyme C and a decapeptide (DTHKSEIAHR) in BSA. Thus, these data will not be shown again. In the present studies, these model proteins were first subjected to "salt-free" transamination before treatment with $BnONH_2$. This treatment was followed by protease digestion using Glu-C/trypsin and LC-MS/MS. Similar to the peptide test, Mascot database searches were conducted using individual protein sequences and a variable modification, i.e. the N-terminal $BnONH_2$ modification (mass shift = +104.03 Da).

As described in section 3.2.1, the N-terminal peptide of lysozyme C (KVFGRCE) in its native state was detected as a doubly charged ion at *m*/*z* = 448.23 (with the carbamidomethylated C residue). The *m*/*z* value of this peptide became 500.24 (*z* = 2) after transamination and the treatment with $BnONH_2$. It corresponded to a mass shift of +104.03 Da, which was equal to the expected value. Figure 4.13 is a composite full mass spectrum that compares the *m*/*z* values between the native and $BnONH_2$-modified N-terminal peptides.

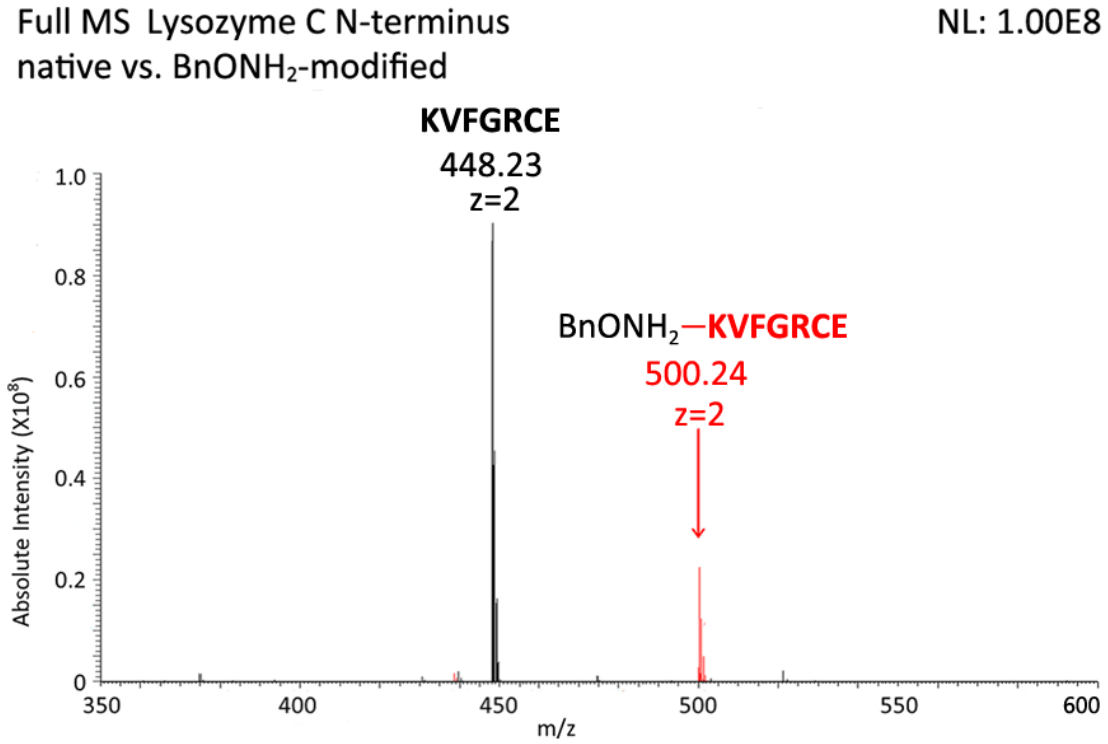


Figure 4.13 Composite full mass spectrum of the N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE) before and after transamination and modification with *O*-benzylhydroxylamine ($BnONH_2$). NL: normalised intensity level; *m*/*z*: mass-to-charge ratio.

The $BnONH_2$-modified N-terminal peptide (*m/z* = 500.24, *z* = 2) was then subjected to HCD fragmentation that generated a series of fragment ions. The resulting MS/MS data were analysed by a Mascot database search. A comparison of the fragment ions showed that the native and $BnONH_2$-modified N-terminal peptides had identical *y*-series ions (Figure 4.14 and Table 4.8). Therefore, the mass shift of +104.03 Da could only be attributed to the intended modifications on the N-terminal K residue. Although *b*-series ions were not matched, these results strongly suggested that transamination and $BnONH_2$ modification took place specifically at the N-terminus of lysozyme C.

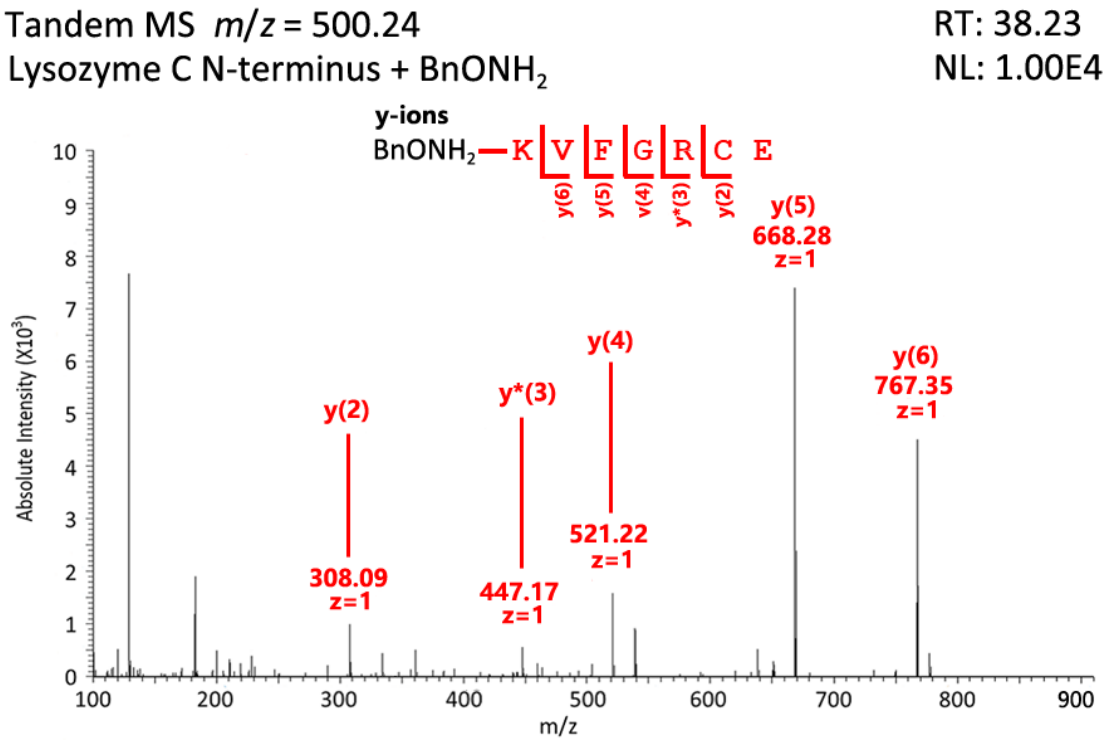


Figure 4.14 Tandem mass spectrum of the $BnONH_2$-modified N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE) with the annotated fragment ions (red). $BnONH_2$: *O*-benzylhydroxylamine; *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio.

Table 4.8 Fragment ions (*in silico* predicted) of the $BnONH_2$-modified N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE). Fragment ions matched to the experimental data are shown in red. $BnONH_2$: *O*-benzylhydroxylamine**.**

#	b	b^{++}	b*	b^{*++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	233.1285	117.0679	216.1019	108.5546	K							7
2	332.1969	166.6021	315.1703	158.0888	V	***767.3505***	384.1789	750.3239	375.6656	749.3399	375.1736	6
3	479.2653	240.1363	462.2387	231.6230	F	***668.2821***	334.6447	651.2555	326.1314	650.2715	325.6394	5
4	536.2867	268.6470	519.2602	260.1337	G	***521.2137***	261.1105	504.1871	252.5972	503.2031	252.1052	4
5	692.3879	346.6976	675.3613	338.1843	R	464.1922	232.5997	447.1656	224.0865	446.1816	223.5945	3
6	852.4185	426.7129	835.3920	418.1996	C	***308.0911***	154.5492			290.0805	145.5439	2
7					E	148.0604	74.5339			130.0499	65.5286	1

As described in section 3.2.2, the tryptic N-terminal peptide of native BSA is composed of 10 amino acid residues: DTHKSEIAHR (with one missed cleavage). Through an LC-MS/MS analysis, this peptide was detected as a triply charged ion at *m*/*z* = 398.54 (Figure 4.15). After transamination and treatment with $BnONH_2$, the modified N-terminal peptide was detected as a triply charged ion at *m*/*z* = 433.21. The *m*/*z* difference between these two ions was equivalent to a mass shift of +104.03 Da, matching the expected value due to $BnONH_2$ modification.

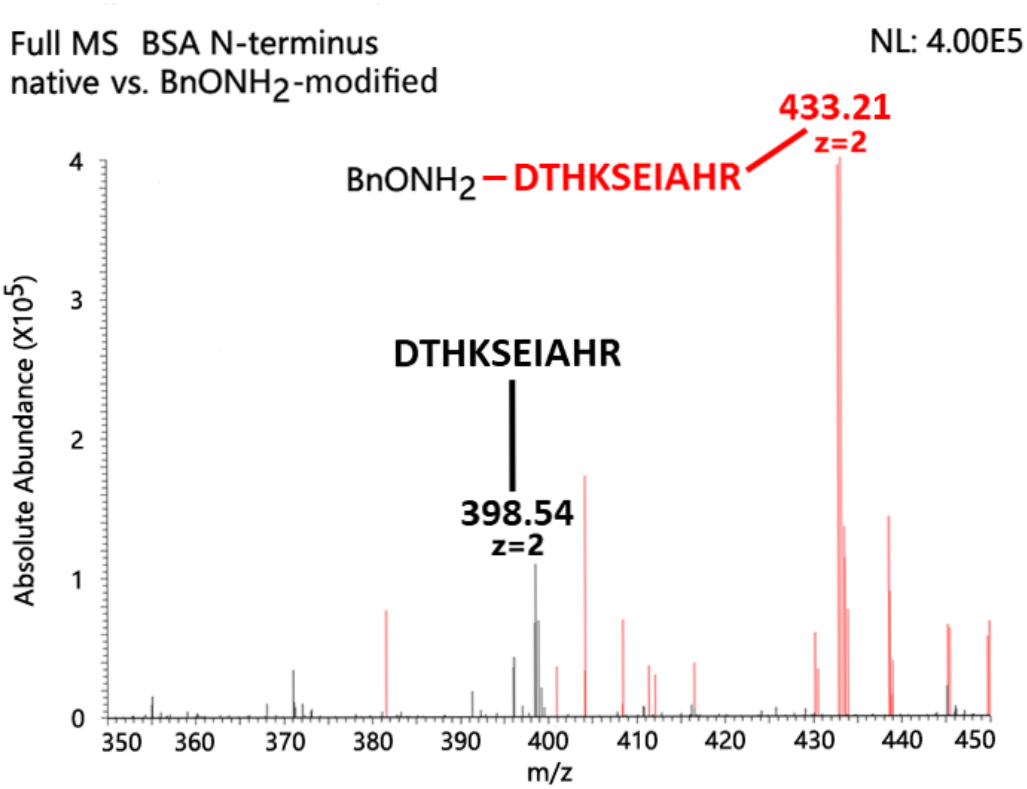


Figure 4.15 Composite full mass spectrum of the N-terminal peptide of BSA (amino acid sequence: DTHKSEIAHR) before and after transamination and *O*-benzylhydroxylamine ($BnONH_2$) modification. NL: normalised intensity level; *m*/*z*: mass-to-charge ratio.

The $BnONH_2$-modified N-terminal peptide was also identified by a Mascot database search. As shown in the corresponding tandem mass spectrum, a series of *y*-ions were detected after HCD fragmentation of the $BnONH_2$-modified N-terminal peptide (4.16 and Table 4.9). These *y*-ions had identical *m*/*z* values with those from the native N-terminal peptide. This result was consistent with the findings in the lysozyme C experiment. In addition, two *y*-ions (y_2 and y_5) that could not be assigned by Mascot were found in the tandem mass spectrum by manual inspection. To summarise, both lysozyme C and BSA experiments showed that $BnONH_2$ modification was specific for the transaminated protein N-termini.

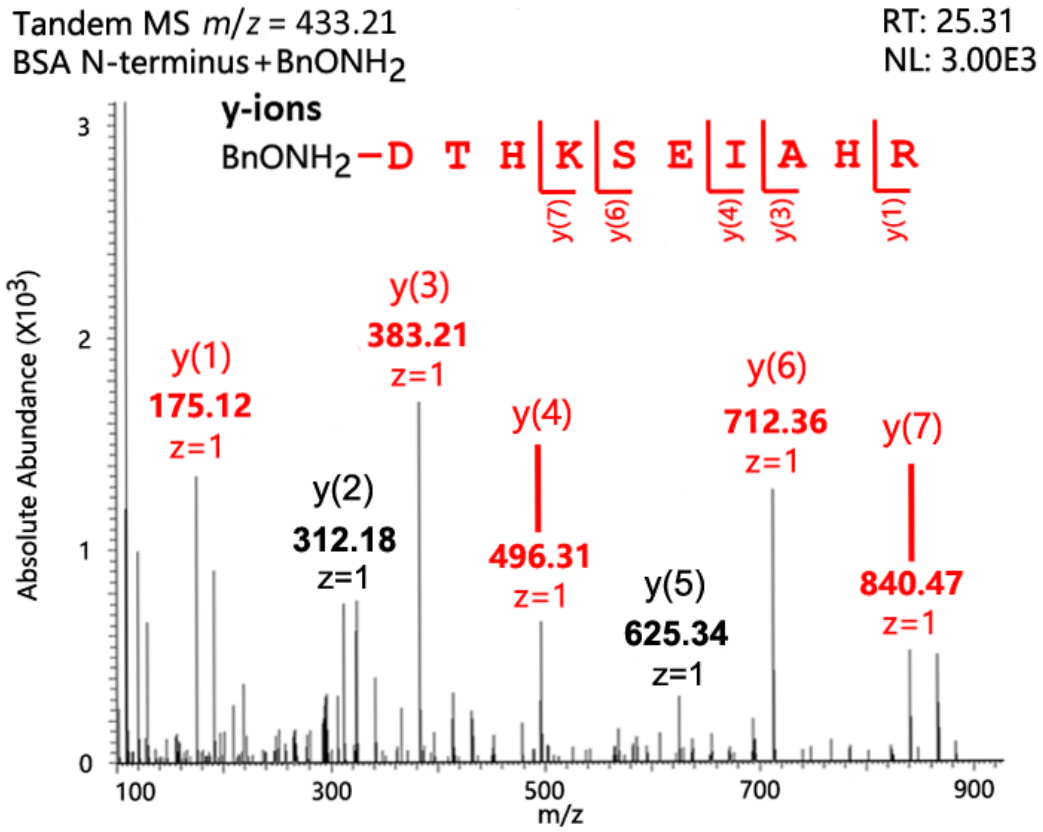


Figure 4.16 Tandem mass spectrum of the $BnONH_2$-modified N-terminal peptide of BSA (amino acid sequence: DTHKSEIAHR). A table of the expected *m/z* values for all possible fragment ions is included below. The annotated ions are labelled in red, whereas putative fragment ions (not annotated by Mascot) are labelled in black. $BnONH_2$: *O*-benzylhydroxylamine; *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio.

Table 4.9 Fragment ions (*in silico* predicted) of the $BnONH_2$-modified N-terminus of BSA (amino acid sequence: DTHKSEIAHR). Fragment ions matched to the experimental data are shown in red. $BnONH_2$: *O*-benzylhydroxylamine.

#	b	b^{++}	b*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	220.0604	110.5339			202.0499	101.5286	D							10
2	321.1081	161.0577			303.0975	152.0524	T	1078.5752	539.7912	1061.5487	531.2780	1060.5647	530.7860	9
3	458.1670	229.5872			440.1565	220.5819	H	*977.5275*	489.2674	960.5010	480.7541	959.5170	480.2621	8
4	586.2620	293.6346	569.2354	285.1214	568.2514	284.6293	K	*840.4686*	420.7380	823.4421	412.2247	822.4581	411.7327	7
5	673.2940	337.1506	656.2675	328.6374	655.2835	328.1454	S	*712.3737*	356.6905	695.3471	348.1772	694.3631	347.6852	6
6	802.3366	401.6719	785.3101	393.1587	784.3260	392.6667	E	625.3416	313.1745	608.3151	304.6612	607.3311	304.1692	5
7	915.4207	458.2140	898.3941	449.7007	897.4101	449.2087	I	*496.2990*	248.6532	479.2725	240.1399			4
8	986.4578	493.7325	969.4312	485.2193	968.4472	484.7272	A	*383.2150*	192.1111	366.1884	183.5979			3
9	1123.5167	562.2620	1106.4902	553.7487	1105.5061	553.2567	H	312.1779	156.5926	295.1513	148.0793			2
10							R	*175.1190*	88.0631	158.0924	79.5498			1

Furthermore, an error-tolerant search also detected $BnONH_2$ modification of the decarboxylated N-terminal peptide of BSA, which exhibited a further mass shift of -44 Da (Table 4.10). A comparison of the number of PSMs between the expected and decarboxylated N-terminal peptides suggested that decarboxylation might dominate over

transamination as far as the N-terminal D residue was concerned. The error-tolerant search also suggested that the $BnONH_2$-modified peptide might be detected in the form of $[M + H + Na]^{2+}$ (further mass shift = +22 Da) or $[M + H + K]^{2+}$ (further mass shift = +38 Da).

Table 4.10 Result of the Mascot error-tolerant for the $BnONH_2$-modified N-terminal peptide of BSA[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	ppm	Score	Duplicate PSM No.
$BnONH_2$-modifed peptide (Reference)					
DTHKSEIAHR (1 – 10) + $BnONH_2$ (N-term)	**433.2122**	**1296.6211**	**-4.75**	**42**	**1**
Putative $BnONH_2$-modifed peptides					
DTHKSEIAHR (1 – 10) + $BnONH_2$ (N-term); -44 Da	**418.5492**	**1252.6313**	**-4.32**	**62**	**3**
DTHKSEIAHR (1 – 10) + $BnONH_2$ (N-term); +22 Da	**660.3050**	**1318.6030**	**-5.70**	**46**	
DTHKSEIAHR (1 – 10) + $BnONH_2$ (N-term); +38 Da	**668.2969**	**1334.5770**	**1.67**	**57**	

[a] A single PSM is shown for each putative $BnONH_2$-modified peptide together with the expected peptide identified in the same experiment as reference. Peptide score ≥ 13 indicates identity. $BnONH_2$: *O*-benzylhydroxylamine; *m/z*: mass-to-charge ratio; MW: molecular weight; ppm: parts per million; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

The second stage of the protein test involved transamination and biotin tagging of lysozyme C and BSA. Initially, a hydrazide-based biotin tag (i.e. hydrazide-biotin) was employed in the test. After treatment of the transaminated proteins with hydrazide-biotin, the expected mass shift would be +239.10 Da (compared to the native peptides). Accordingly, MS/MS data of the biotin-treated lysozyme C and BSA samples were subjected to Mascot searches against their respective protein sequences, with Nt-biotinylation (+239.10 Da) set as one of the variable modifications. However, the Mascot database searches did not identify any biotinylated N-terminal peptide in either the lysozyme C or BSA samples despite the repeated attempts (N = 4) to locate the biotin tag.

Error-tolerant searches were thus performed on the MS/MS data of either protein sample. In the case of lysozyme C, the error-tolerant search identified the putative biotinylated N-terminal peptide with further mass shifts (Table 4.11). Among them, the further mass shifts of +43 and +57 Da were attributed to carbamylation (K) and carbamidomethylation (K), respectively. Such assignments might be genuine since the protein samples were indeed dissolved in urea and treated with iodoacetamide, which would lead to the said

modifications (see section 4.3). Other putative modifications included formylation (mass shift = +28 Da) and a complete loss of the K side chain (N-terminal K to G, mass shift = -71 Da). Although such assignments remain to be verified, formylation has been reported as an experimental artefact due to prolonged exposure to formic acid (Zheng and Doucette, 2016). On the other hand, the N-terminal peptide of BSA was still not identified. Therefore, the error-tolerant searches did not provide definitive assignments of the N-terminal biotin tag.

Table 4.11 Result of the Mascot error-tolerant search for the N-terminal peptide of transaminated lysozyme C, which was further treated with hydrazide-biotin[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	ppm	Score	Duplicate PSM No.
Native peptide (Reference)					
KVFGRCE (1 – 10) + Carbamidomethyl (C)	448.2266	894.4382	0.44	44	31
Putative biotinylated peptides					
KVFGRCE (1 – 10) + Carbamidomethyl (C); Biotin (N-term); -71 Da	532.2249	1062.4375	-2.13	21	
KVFGRCE (1 – 10) + Carbamidomethyl (C); Biotin (N-term); +28 Da	581.7587	1161.5059	-2.57	20	5
KVFGRCE (1 – 10) + Carbamidomethyl (C); Biotin (N-term); +43 Da	589.2644	1176.5168	-2.14	24	
KVFGRCE (1 – 10) + Carbamidomethyl (C); Biotin (N-term); +57 Da	596.2726	1190.5325	0.39	22	9

[a] Orbitrap RAW data of four replicate analyses were pooled and searched in Mascot twice, with cysteine (C) carbamidomethylation set as either a fixed or variable modification. Both Mascot searches produced the same result, which was edited to show a single PSM for each putative biotinylated peptide. The native peptide confidently identified in the same experiment is also included as reference. Peptide score ≥ 13 indicates identity. *m/z*: mass-to-charge ratio; MW: molecular weight; ppm: parts per million; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

In response to these negative results, another analytical method was employed to test the same set of samples at a different level. Dot blotting is a simple immunoblotting technique that detects proteins of interest using specific antibodies. Since this method does not require protein separation by electrophoresis, the time of analysis can be dramatically reduced. It is especially suitable for analysing single protein samples. In the present experiments, the model protein samples yielding negative results were re-analysed by dot blotting to check for the presence of a biotin signal. The protein samples of lysozyme C or BSA, which had been subjected to transamination and subsequently to biotin tagging, were analysed together with three negative controls: native protein, transamination-only, and biotinylation-only.

As shown in Figure 4.17, the results with the lysozyme C and BSA samples indicated that the biotin tag was successfully attached to both proteins after transamination and the treatment with hydrazide-biotin. The biotin signal was absent in either the native protein (Trans-, Biotin-) or transamination-only (Trans+, Biotin-) samples. In contrast, a strong biotin signal was detected in the treated samples (Trans+, Biotin+). Interestingly, a biotin signal was also detected in the biotinylation-only samples (Trans-, Biotin+), albeit to a much lesser extent. Given the equal loading across all four samples confirmed by Ponceau S staining, these results demonstrated that transamination introduced more carbonyl groups for hydrazide-biotin to react with. The biotin signal in one of the negative controls (i.e. biotinylation-only) might reflect spontaneous protein carbonylation (see section 4.3).

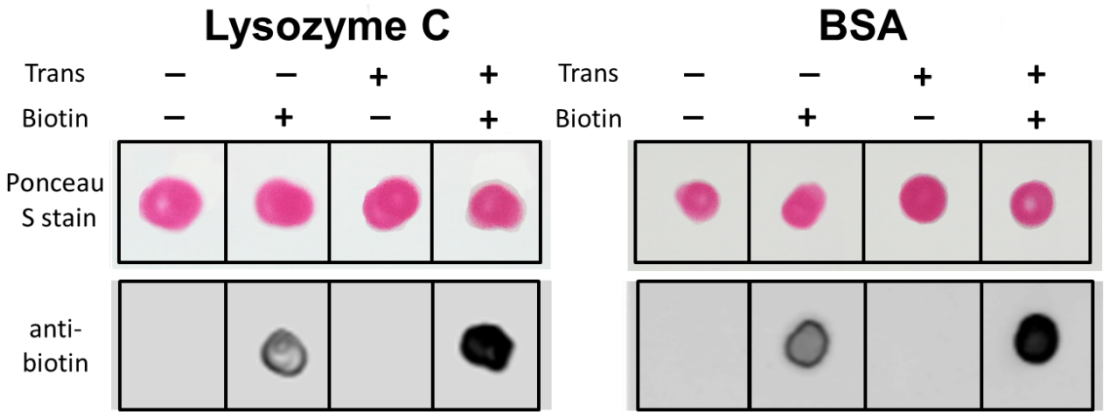


Figure 4.17 Dot blotting of lysozyme C and BSA, which were subjected to selective transamination and treatment with hydrazide-biotin. For each protein, three negative controls and the treated sample (from left to right) are: native protein, biotinylation-only, transamination-only, and both transamination and biotinylation. Protein loading was checked by Ponceau S staining before incubation with the biotin-specific antibody. Trans: transamination.

Having confirmed that the biotin tag was successfully added to the model proteins after transamination, the next step in this test was to confirm the biotin tagging of protein N-termini by LC-MS/MS. Since the previous experiments failed to do so using hydrazide-biotin, alkoxyamine-PEG_4-biotin was employed instead. This reagent specifically attaches biotin to a carbonyl group by forming an oxime bond instead of a hydrazone bond (Figure 4.18). The experimental procedure was the same as previously described where the protein samples were first analysed by dot blotting and then digested with a protease for LC-MS/MS.

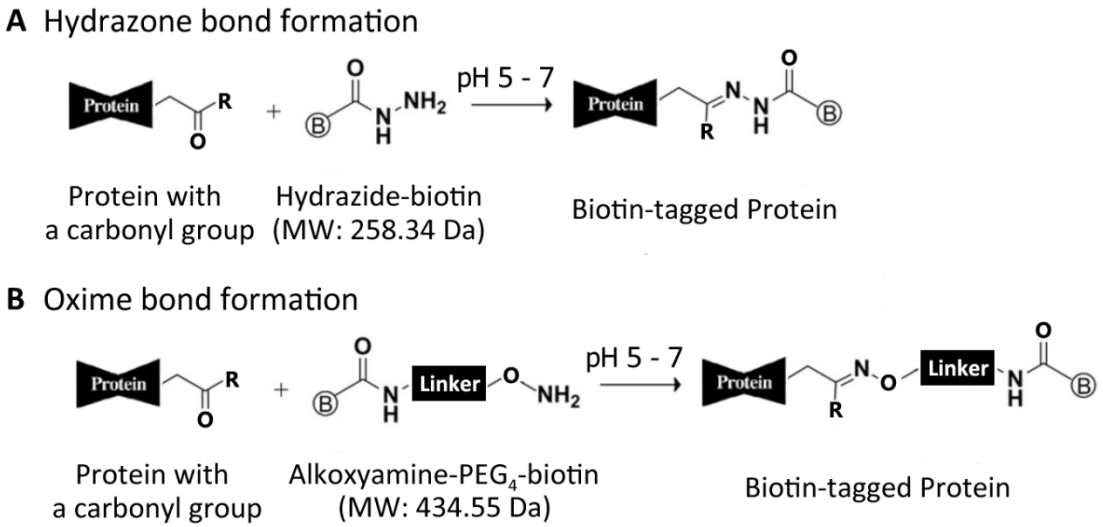


Figure 4.18 Illustrated mechanism of the carbonyl-specific biotinylation at protein N-terminus. The N-terminal carbonyl group can be introduced by selective transamination. (**A**) Biotinylation of the carbonyl group through hydrazone bond formation. (**B**) Biotinylation of the carbonyl group through oxime bond formation. The linker region in alkoxyamine-PEG_4-biotin is not shown. Ⓑ: biotin; R: side chain of the N-terminal amino acid residue.

As shown in Figure 4.19, the dot blotting analyses revealed that a strong biotin signal was detected in the treated samples (Trans+, Biotin+) of both lysozyme C and BSA. In contrast, the biotin signal was absent in two negative control samples: native protein (Trans-, Biotin-) and transamination-only (Trans+, Biotin-). Similar to the results with hydrazide-biotin, a weak signal of biotin was also detected in the biotinylation-only (Trans-, Biotin+) samples of both lysozyme C and BSA.

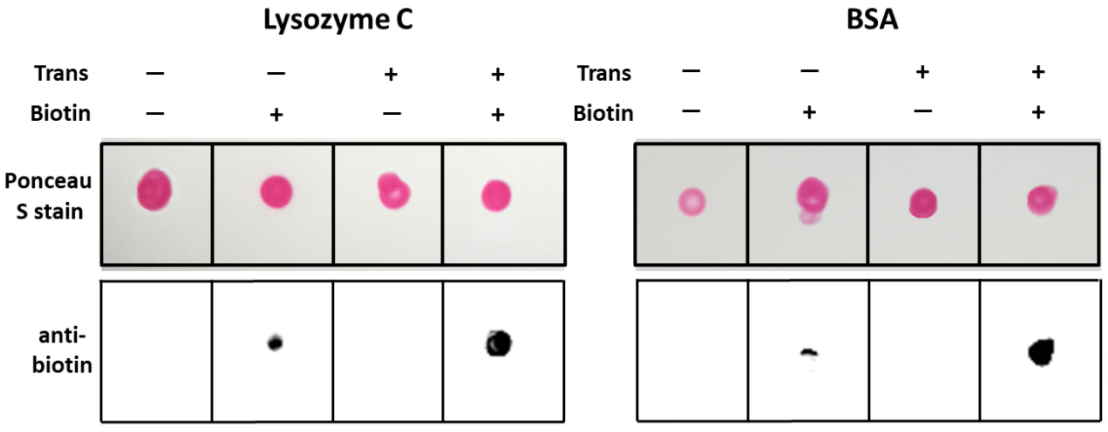


Figure 4.19 Dot blotting analyses of lysozyme C and BSA tagged with alkoxyamine-PEG_4-biotin after selective transamination. For each protein, three negative controls and the treated sample (from left to right) are: native protein, biotinylation-only, transamination-only, and both transamination and biotinylation. Protein loading was checked by Ponceau S staining before incubation with the biotin-specific antibody. Trans: transamination.

In regard to the LC-MS/MS analysis, a Mascot database search identified the biotinylated N-termini of both lysozyme C and BSA. For lysozyme C, the biotinylated N-terminal peptide was detected in the full mass spectrum (*RT* = 36.65 min) as a doubly charged ion at *m*/*z* = 655.81 (Figure 4.20). The mass difference between the biotinylated peptide and its native counterpart was calculated to be 415.17 Da, matching the expected mass shift due to biotinylation. In addition, inspection of the corresponding tandem mass spectrum localised the biotin tag to the N-terminus of the peptide (Figure 4.21 and Table 4.12). It should be noted that, although Nt-biotinylation and biotinylation (K) were indistinguishable in this case, double biotinylation of the N-terminal peptide (at both the N-terminus and the K residue) was never detected when the Mascot search was repeated with biotinylation (K) set as a variable modification.

Consistent with the findings in the previous test with human ACTH (Figure 4.12), the fragment ion that likely signifies the biotin tag cleavage (*m*/*z* = 270.13, *z* = 1) was once again present with high intensity. It was complemented by two singly charged ions (*m*/*z* = 876.40 or 1041.50), which were putative remnants (left by the biotin tag cleavage) of the b^*_6 fragment ion and the biotinylated peptide itself, respectively. Nevertheless, there were fragment ions (e.g. *m*/*z* = 435.23, *z* = 1) that could not be attributed to the biotin tag cleavage and hence remained to be elucidated.

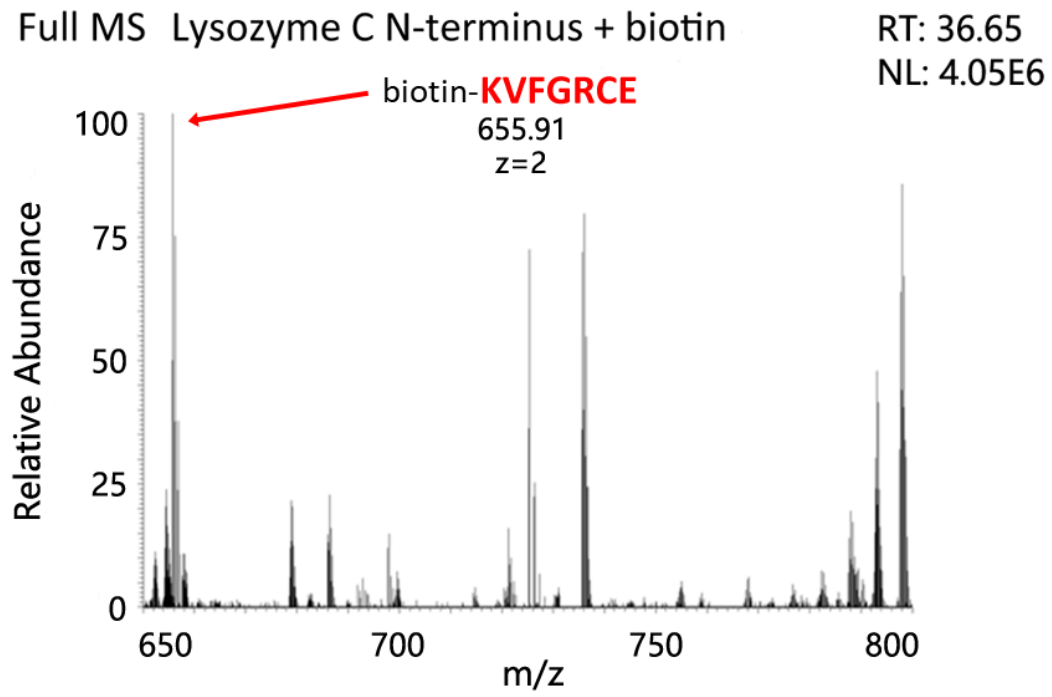


Figure 4.20 Full mass spectrum of the biotinylated N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE), which is shown as a doubly charged ion at *m*/*z* = 655.91. *RT*: retention time; NL: normalise intensity level; *m*/*z*: mass-to-charge ratio.

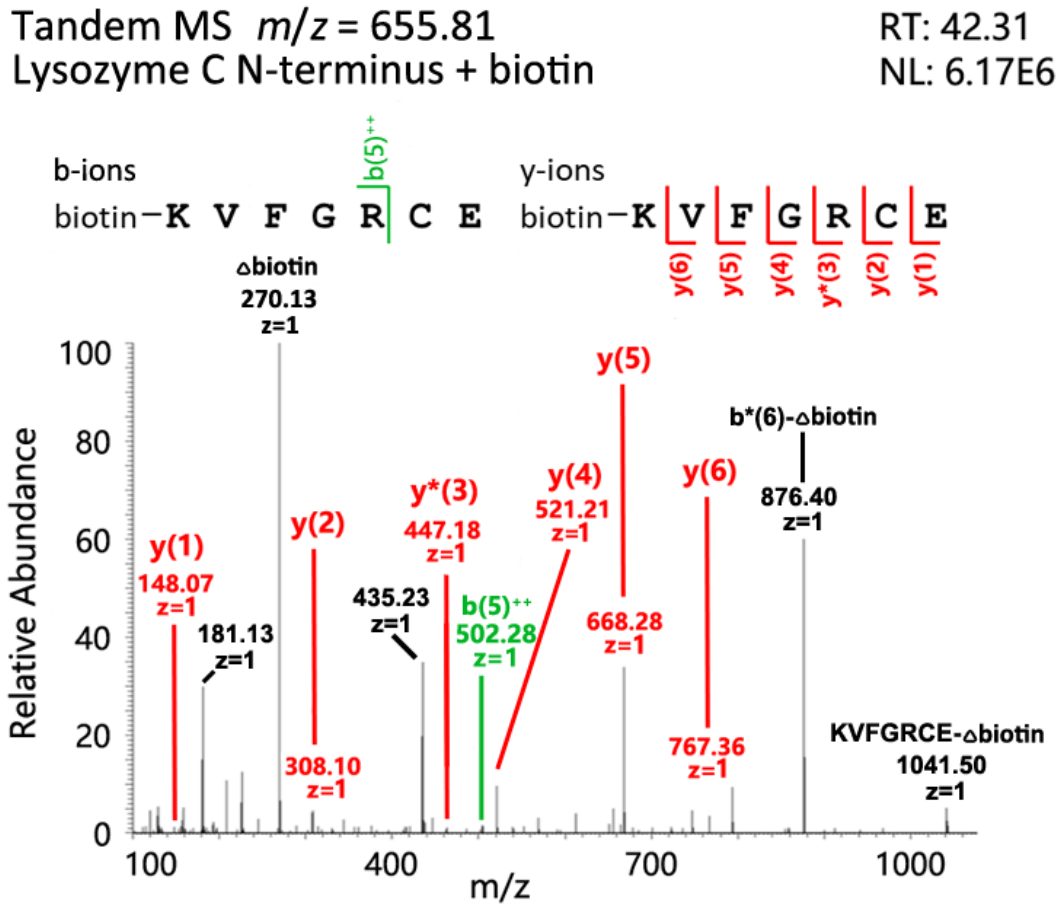


Figure 4.21 Tandem mass spectrum of the biotinylated N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE). A table of the expected *m/z* values for all possible fragment ions is shown in Table 4.12. The annotated *b*- and *y*-ions are labelled in green and red, respectively. Putative fragment ions (not annotated by Mascot) are labelled in black, including the singly charged ion at *m/z* = 270.13 that likely resulted from putative cleavage of the attached biotin tag (denoted by $_{\Delta}$biotin). *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio.

Table 4.12 Fragment ions (*in silico* predicted) of the biotinylated N-terminal peptide of lysozyme C (amino acid sequence: KVFGRCE). Fragment ions matched to the experimental data are shown in red.

#	b	b^{++}	b*	b^{*++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	544.2799	272.6436	527.2534	264.1303	K							7
2	643.3484	322.1778	626.3218	313.6645	V	*767.3505*	384.1789	750.3239	375.6656	749.3399	375.1736	6
3	790.4168	395.7120	773.3902	387.1988	F	*668.2821*	334.6447	651.2555	326.1314	650.2715	325.6394	5
4	847.4382	424.2228	830.4117	415.7095	G	*521.2137*	261.1105	504.1871	252.5972	503.2031	252.1052	4
5	1003.5393	502.2733	986.5128	493.7600	R	*464.1922*	232.5997	447.1656	224.0865	446.1816	223.5945	3
6	1163.5700	582.2886	1146.5434	573.7754	C	*308.0911*	154.5492			290.0805	145.5439	2
7					E	*148.0604*	74.5339			130.0499	65.5286	1

Similarly, the biotin-tagged N-terminal peptide of BSA was eluted at *RT* = 21.87 min and detected in the corresponding full mass spectrum as a doubly charged ion at *m*/*z* = 804.89 (Figure 4.22). The calculated mass shift was equal to the expected value (415.17 Da), thus confirming the addition of the biotin tag. Although *b*-series fragment ions were not directly detected, the MS/MS data strongly suggested that this biotin tag was specifically attached to the N-terminal residue of this peptide (Figure 4.23 and Table 4.13). In keeping with the previous results, the intense peak at *m*/*z* = 270.13 (*z* = 1) was also detected. It likely signifies the cleavage of the attached biotin tag during HCD fragmentation (discussed in section 4.3). Furthermore, biotinylation (K) of the N-terminal peptide could not be detected when the Mascot search was repeated to include it in the list of variable modifications.

It should be noted that the decarboxylated N-terminal peptide (mass shift = -44 Da) was likely also tagged with alkoxyamine-PEG_4-biotin. The putative decarboxylated and biotinylated N-terminal peptide was identified as a doubly charged ion at *m*/*z* = 782.89, corresponding to a MW of 1563.78 Da. In conclusion, the results of both the lysozyme C and BSA tests indicated that transamination enabled the biotin tagging of the modified protein N-termini through oxime bond formation.

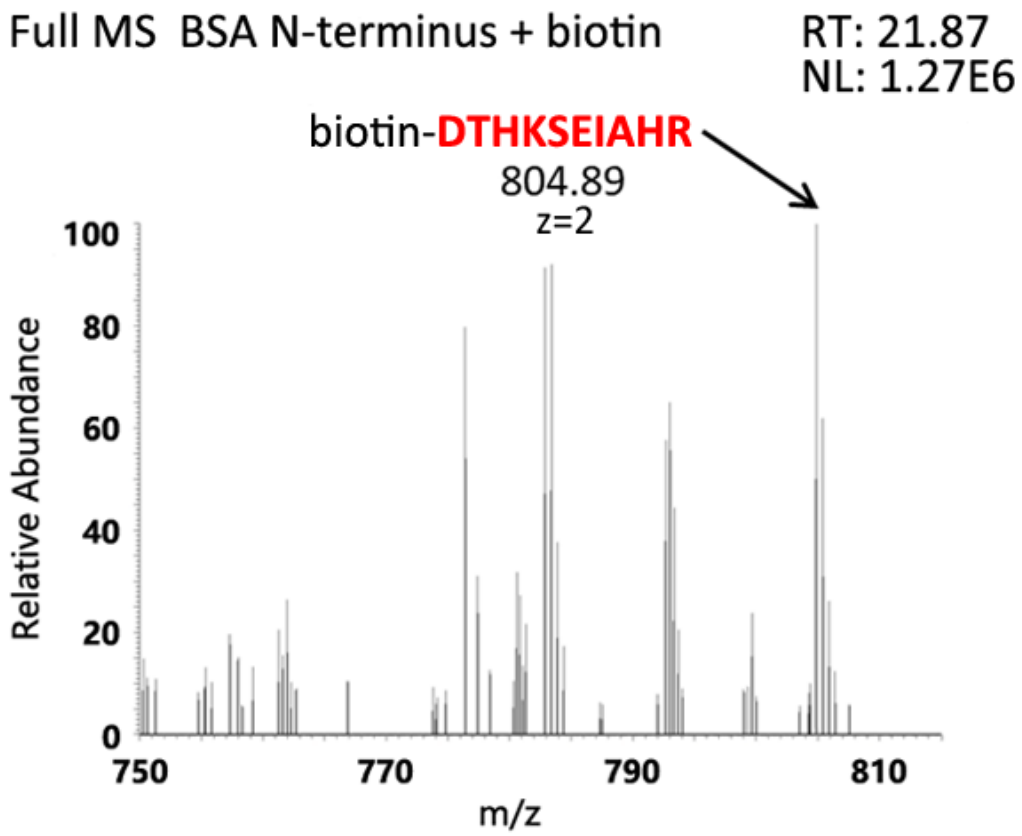


Figure 4.22 Full mass spectrum of the biotinylated N-terminal peptide of BSA (amino acid sequence: DTHKSEIAHR), which was detected as a doubly charged ion at *m*/*z* = 804.89. *RT*: retention time; NL: normalised intensity level; *m*/*z*: mass-to-charge ratio.

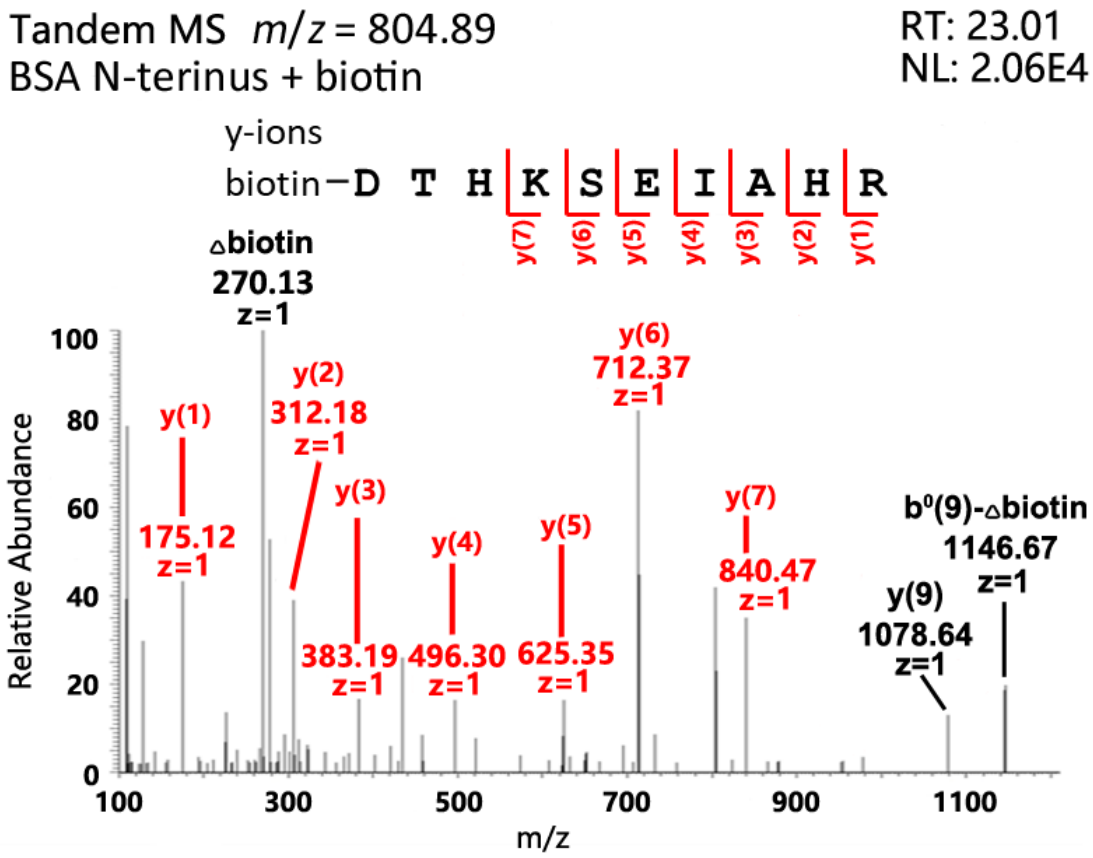


Figure 4.23 Tandem mass spectrum of the biotinylated N-terminal peptide of BSA (amino acid sequence: DTHKSEIAHR). A table of the expected *m/z* values for all possible fragment ions is shown in Table 4.13. The annotated ions are labelled in red, whereas putative fragment ions (not annotated by Mascot) are labelled in black. In particular, the singly charged ion at *m/z* = 270.13, which likely resulted from putative cleavage of the attached biotin tag, is denoted by $_{\Delta}$biotin. *RT*: retention time; NL: normalised intensity level; *m/z*: mass-to-charge ratio.

Table 4.13 Fragment ions (*in silico* predicted) of the biotinylated N-terminal peptide of BSA (amino acid sequence: DTHKSEIAHR). Fragment ions matched to the experimental data are shown in red.

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	531.2119	266.1096			513.2014	257.1043	D							10
2	632.2596	316.6334			614.2490	307.6282	T	1078.5752	539.7912	1061.5487	531.2780	1060.5647	530.7860	9
3	769.3185	385.1629			751.3080	376.1576	H	977.5275	489.2674	960.5010	480.7541	959.5170	480.2621	8
4	897.4135	449.2104	880.3869	440.6971	879.4029	440.2051	K	*840.4686*	420.7380	823.4421	412.2247	822.4581	411.7327	7
5	984.4455	492.7264	967.4190	484.2131	966.4349	483.7211	S	*712.3737*	356.6905	695.3471	348.1772	694.3631	347.6852	6
6	1113.4881	557.2477	1096.4616	548.7344	1095.4775	548.2424	E	*625.3416*	313.1745	608.3151	304.6612	607.3311	304.1692	5
7	1226.5722	613.7897	1209.5456	605.2764	1208.5616	604.7844	I	*496.2990*	248.6532	479.2725	240.1399			4
8	1297.6093	649.3083	1280.5827	640.7950	1279.5987	640.3030	A	*383.2150*	192.1111	366.1884	183.5979			3
9	1434.6682	717.8377	1417.6416	709.3245	1416.6576	708.8325	H	*312.1779*	156.5926	295.1513	148.0793			2
10							R	*175.1190*	88.0631	158.0924	79.5498			1

The next major step of the present study was to enrich the biotin-tagged proteins by AP using immobilised avidin resins. Two types of such resins were employed in this test: NeutrAvidin agarose and monomeric avidin agarose. As described previously, NeutrAvidin is a deglycosylated form of the avidin protein (MW: 60 kDa), which consists of four identical subunits (MW: 15 kDa) each binding to a single biotin. Deglycosylation is claimed to reduce nonspecific binding (Marttila *et al.*, 2000). In terms of the affinity for biotin, the dissociation constant (K_d) of NeutrAvidin is 10^{-15} M. On the other hand, monomeric avidin has a much lower affinity for biotin (K_d = 10^{-8} M), allowing for mild elution of the biotinylated molecules and recycling of this reagent (Henrikson *et al.*, 1979).

In the AP tests, the biotin-tagged BSA was incubated with NeutrAvidin agarose first. After extensive washing, the NeutrAvidin beads were boiled in SDS sample buffer to elute the bound proteins for SDS-PAGE analysis. Two sets of BSA samples (eight in total) were separated on a pair of 10 % polyacrylamide gels. The first gel consisted of the test sample (transamination and biotinylation) and three negative controls (native protein, transamination-only, and biotinylation-only) before AP, whereas the second gel contained the four equivalent samples after AP. Equal amounts of BSA were detected in all four samples before AP (Figure 4.24, left). The major protein band (MW: ~ 66 kDa) corresponded to BSA, whereas the additional bands were routinely detected impurities.

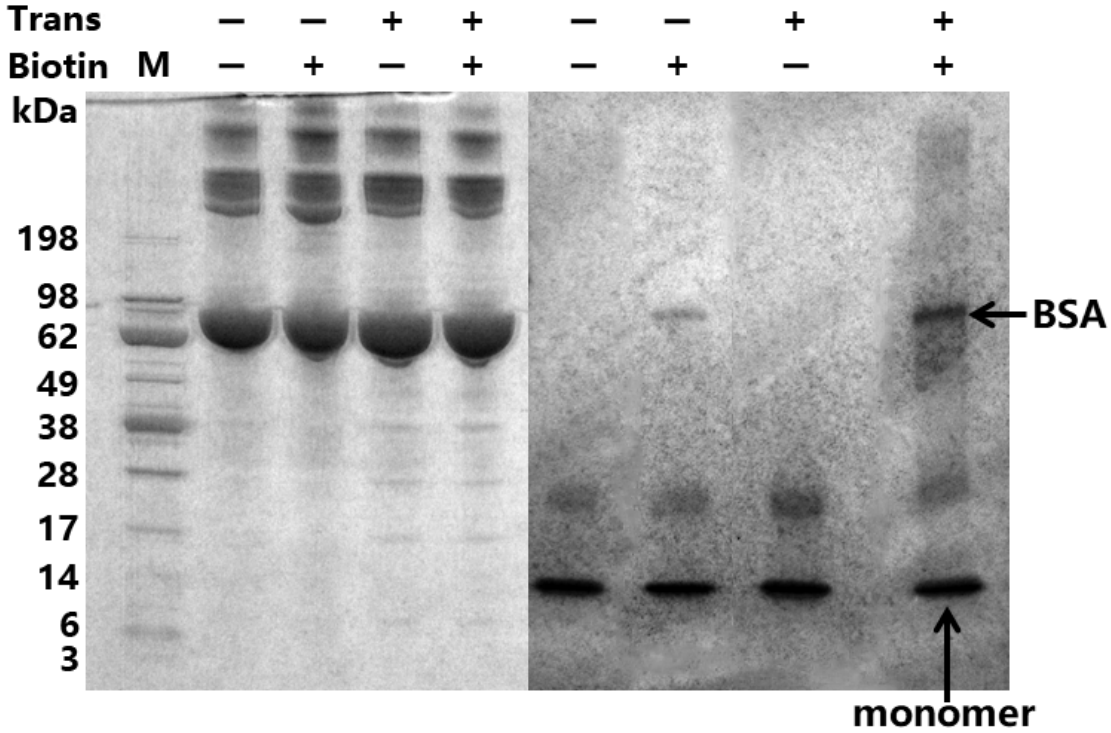


Figure 4.24 Sodium dodecyl sulfate–polyacrylamide gel electrophoresis **(**SDS-PAGE) of the biotinylated BSA before (left) and after (right) affinity purification with NeutrAvidin agarose. The position of NeutrAvidin monomers is also indicated. Trans: transamination; M: protein marker.

As shown in Figure 4.24 (right), the BSA band was mainly detected in the treated sample (Trans+, Biotin+) after AP. This result suggested that a biotin tag had been successfully attached to the transaminated BSA and that NeutrAvidin agarose was able to enrich the biotin-tagged protein. In keeping with the dot blotting results, the BSA band was also detected in the biotinylation-only sample (Trans-, Biotin+) but with lower intensity. This observation indicated that biotinylation could also take place spontaneously (i.e. without transamination), albeit to a lesser extent. Importantly, two additional bands were also detected in all four samples after AP. The major band of the two corresponded to a MW of 15 kDa. A plausible explanation for this observation is that the harsh conditions for elution damaged the NeutrAvidin proteins, causing dissociation of the tetramers into mainly monomers (MW: 15 kDa) and possibly dimers (the minor band; MW: 30 kDa) too.

Since the MW of NeutrAvidin monomers is similar to that of lysozyme C (14 kDa), it would be difficult to separate these two proteins by SDS-PAGE. NeutrAvidin agarose was thus replaced by monomeric avidin agarose to enrich for the biotinylated lysozyme C. Similarly, the result of SDS-PAGE showed that the protein band of lysozyme C was present in equal quantities across all the samples before AP but only detected in the treated sample (Trans+, Biotin+) after AP. In contrast, none of the control samples exhibited any detectable band after AP (Figure 4.25). This result demonstrated that monomeric avidin agarose was a viable option for AP of the biotinylated proteins owing to its desired specificity and the mild elution conditions, the latter of which eliminated potential inference from avidin monomers.

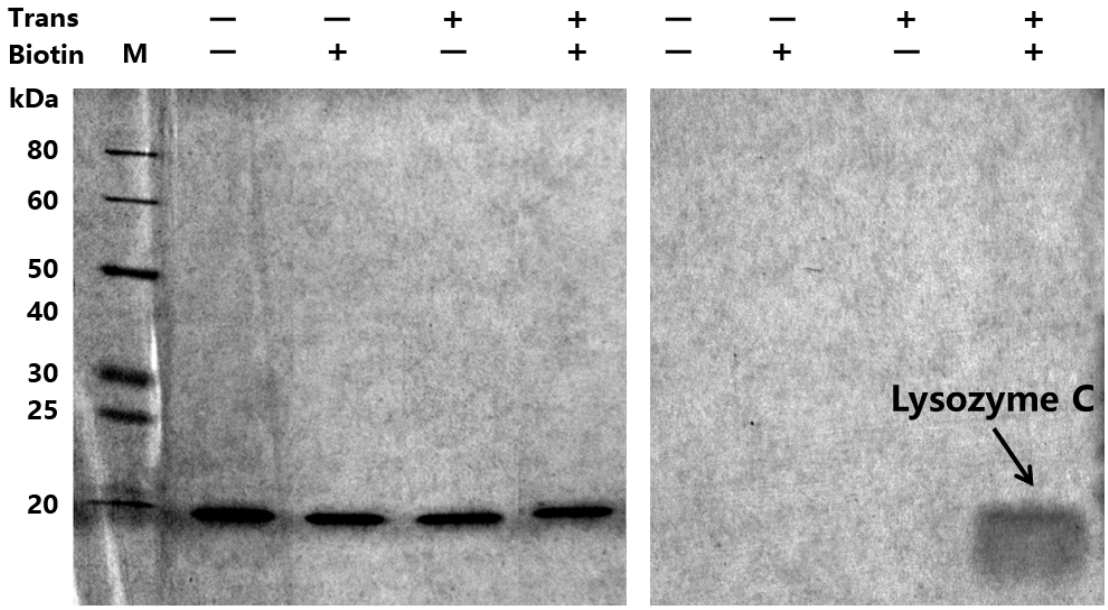


Figure 4.25 Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) of the biotinylated lysozyme C before (left) and after (right) affinity purification with monomeric avidin agarose. Trans: transamination; M: protein markers.

Following the success of AP at the protein level, monomeric avidin agarose was also employed to enrich the biotinylated N-terminal peptides of lysozyme C and BSA after protease digestion. However, the biotinylated N-terminal peptide was not detected by LC-MS/MS after AP for either protein. Instead, the native N-terminal peptide of lysozyme C and a number of internal peptides were still detected, as reflected by the protein sequence coverage (Figure 4.26). These results might point to the difficulties in completely removing internal peptides even using a positive selection strategy.

Lysozyme C

Protein sequence coverage: 24%

Matched peptides shown in ***bold red***.

```
  1 KVFGRCELAA AMKRHGLDNY RGYSLGNWVC AAKFESNFNT QATNRNTDGS
 51 TDYGILQINS RWWCNDGRTP GSRNLCNIPC SALLSSDITA SVNCAKKIVS
101 DGNGMNAWVA WRNRCKGTDV QAWIRGCRL
```

BSA

Protein sequence coverage: 42%

Matched peptides shown in ***bold red***.

```
  1 DTHKSEIAHR FKDLGEEHFK GLVLIAFSQY LQQCPFDEHV KLVNELTEFA
 51 KTCVADESHA GCEKSLHTLF GDELCKVASL RETYGDMADC CEKQEPERNE
101 CFLSHKDDSP DLPKLKPDPN TLCDEFKADE KKFWGKYLYE IARRHPYFYA
151 PELLYYANKY NGVFQECCQA EDKGACLLPK IETMREKVLA SSARQRLRCA
201 SIQKFGERAL KAWSVARLSQ KFPKAEFVEV TKLVTDLTKV HKECCHGDLL
251 ECADDRADLA KYICDNQDTI SSKLKECCDK PLLEKSHCIA EVEKDAIPEN
301 LPPLTADFAE DKDVCKNYQE AKDAFLGSFL YEYSRRHPEY AVSVLLRLAK
351 EYEATLEECC AKDDPHACYS TVFDKLKHLV DEPQNLIKQN CDQFEKLGEY
401 GFQNALIVRY TRKVPQVSTP TLVEVSRSLG KVGTRCCTKP ESERMPCTED
451 YLSLILNRLC VLHEKTPVSE KVTKCCTESL VNRRPCFSAL TPDETYVPKA
501 FDEKLFTFHA DICTLPDTEK QIKKQTALVE LLKHKPKATE EQLKTVMENF
551 VAFVDKCCAA DDKEACFAVE GPKLVVSTQT ALA
```

Figure 4.26 Sequence coverage of the putative biotinylated lysozyme C (upper) and BSA (lower) after affinity purification with monomeric avidin agarose at the peptide level.

It should also be noted that a plethora of confounding factors can influence the detection/non-detection of a peptide ion, including its abundance, ionisation/fragmentation efficiency, and so on (Li *et al.*, 2010). These confounding factors manifest as the peptide "flyability" (a generic term used to describe the intrinsic detectability of the peptide), which has recently been mathematically interpreted using absolute quantitation data of the yeast proteome (Jarnuczak *et al.*, 2016). Therefore, the contaminating internal peptides might have higher "flyability" and thus dominated the LC-MS/MS results even if they had been reduced in abundance after AP. Possibly due to the lower "flyability", the biotinylated N-terminal peptides might have been out-competed by the internal peptides for ionisation even if they had been enriched by AP, which in principle also explains the absence of such peptides in the LC-MS/MS results.

In response to these negative results, error-tolerant searches were performed again to identify any biotinylated N-terminal peptide in the enriched samples. Potentially such peptides were undetectable due to unsuspected further modifications. Although there were no reliable search hits for the biotinylated N-terminal peptide of lysozyme C, the putative biotinylated N-terminal peptide was indeed identified in the BSA dataset (Table 4.14). Consistent with the previous results, a mass shift of -44 Da was fitted to the N-terminal peptide on top of the intended modification. It strongly supported the conjecture that the N-terminal D residue of a protein or peptide was decarboxylated after transamination and could receive a carbonyl-reactive biotin tag. In addition, a further mass shift of -28 Da was also detected, likely indicating a partial loss of the D residue side chain. The outcome of this unsuspected modification was the conversion from a carboxyl to a hydroxyl group (i.e. N-terminal D to S).

Table 4.14 Result of the Mascot error-tolerant search for the biotinylated N-terminal peptide of BSA after affinity purification[a].

Peptide-Spectrum Match (Start – End)	*m/z*	MW	ppm	Score	Duplicate PSM No.
Internal peptides (Reference)					
LVNELTEFAK (42 – 51)	**582.3142**	**1162.6234**	**-8.27**	**51**	
CASIQKFGER (199 – 208) + Carbamidomethyl (C)	**598.2927**	**1194.5815**	**-9.01**	**50**	**2**
Putative biotinylated N-terminal peptides					
DTHKSEIAHR (1 – 10) + Biotin (N-term); -44 Da	**782.8916**	**1563.7828**	**-8.97**	**49**	**8**
DTHKSEIAHR (1 – 10) + Biotin (N-term); -28 Da	**527.5950**	**1579.7777**	**-9.27**	**20**	

[a] Orbitrap RAW data were processed and searched in Mascot twice, with cysteine (C) carbamidomethylation set as either a fixed or variable modification. Both Mascot searches showed the same result, which was edited to show a single PSM for each putative biotinylated N-terminal peptide. Two internal peptides that were confidently identified in the same experiment are also included as reference. Peptide score ≥ 13 indicates identity. *m/z*: mass-to-charge ratio; MW: molecular weight; ppm: parts per million; Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM).

4.3 Discussion

In addition to negative selection strategies such as NHS-Sepharose, N-terminalomics also employs positive selection strategies to retain peptides that can shed light on the identity of true protein N-termini. Similar to the Subtiligase and *O*-methylisourea methods that fall in this category, selective transamination also aimed to achieve the enrichment and identification of protein N-termini by exploiting the semi-unique reactivity of protein α-amino groups. In principle, this reaction presents clear advantages over the other methods. Compared to N-CLAP and *O*-methylisourea, selective transamination modifies free protein N-termini in a single step and does not target ε-amino groups on K side chains. Contrary to the Subtiligase method, selective transamination employs a simple combination of small molecules instead of a proprietary enzyme.

To our knowledge, selective transamination has not been utilised at the proteome level to study protein N-termini (or for other purposes). A positive selection strategy was thus proposed by our group in view of the initial reports from Sonomura *et al.* (2009a) that this reaction could be employed for negative selection of the N-terminal peptides of several model proteins. To achieve the positive selection of N-terminal peptides, selective transamination was performed in conjugation with carbonyl-specific biotinylation and AP. Ultimately, this approach can be used to enrich and identify both free protein N-termini and protease cleavage products.

As illustrated in Figure 4.2, the proposed scheme of this approach consisted of five major stages: I. conversion of N-terminal α-amino groups to carbonyl groups through selective transamination; II. attachment of biotin affinity tags to the resulting carbonyl groups using carbonyl-specific biotin reagents; III. digestion of the biotinylated proteins into peptide mixtures using appropriate proteases; IV. enrichment of the biotinylated peptides using immobilised avidin resins; V. identification of protein N-termini by LC-MS/MS.

The proposed approach shared several similarities with other positive selection strategies. For instance, all the positive selection strategies take advantage of the biotin-avidin interactions to enrich for N-terminal peptides, despite that biotinylation is achieved through different chemistries. Ideally, selective transamination should be directly followed by AP that exploits the newly introduced carbonyl group. However, this is currently limited by the choice of carbonyl-specific reagents. As an immobilised carbonyl scavenger, tosylhydrazine glass represents one of the few options for direct capture of peptides with carbonyl groups (Sonomura *et al.*, 2011). But this reagent is not easily accessible and can only be used in a

negative selection fashion (i.e. permanent depletion of peptides with carbonyl groups). In contrast, biotin tags are widely used for AP because of the highly specific and tight binding between biotin and (strept)avidin, as well as the well-defined measures to elute the captured proteins/peptides afterwards. Furthermore, biotin tags are readily available in several carbonyl-specific forms. Accordingly, immobilised avidin/streptavidin beads are used for AP in all positive selection strategies, including our proposed scheme.

To evaluate the feasibility of selective transamination for tagging protein N-termini and AP, systematic tests were conducted at two different levels – peptide and protein. First, synthetic peptides were employed to characterise the reaction products of transamination in a timely fashion. In addition, the peptide test also served to analyse the products of further carbonyl modifications, including biotin tagging. The peptide test paved the way for the subsequent experiments using individual model proteins, which were divided into two parts. The first part of the protein test employed dot blotting and SDS-PAGE to assess the biotin tagging and AP of intact proteins. The second part was equivalent to the peptide test, but the synthetic peptides were replaced by a mixture of proteolytic peptides from the digested model proteins. LC-MS/MS was the primary technique to assess the biotin tagging of protein N-termini and AP.

It should be noted that the order in which the experiment results were presented in this chapter does not necessarily reflect the chronological order in which the experiments were carried out. In the current study, the peptide test was conducted in response to the failure to identify the N-termini of model proteins that were tagged with hydrazide-biotin. Initially, this biotin reagent was preferred due to the simplicity and low cost. However, the biotinylated protein N-termini could not be identified by standard Mascot database searches after the treatment with hydrazide-biotin. As a result, this biotin tag was replaced by two alkoxyamine-based reagents. A thiol-cleavable biotin tag (alkoxyamine-PEG_4-SS-PEG_4-biotin) was employed for the peptide test but then deemed unsuitable for the protein test due to the following reason: this reagent is incompatible with the full denaturation of proteins, which requires the use of DL-dithiothreitol (DTT) to reduce disulfide bonds; DTT can also cleave off the biotin tag by reducing the thiol-cleavable linker. Consequently, alkoxyamine-PEG_4-biotin (without the thiol-cleavable linker) was used instead for the protein test.

Human ACTH (amino acid sequence: SYSMEHFRWG; MW: 1298.55 Da) and rat renin substrate (DRVYIHPFHLLYYS; MW: 1821.92 Da) are two synthetic peptides employed in the peptide test. According to the base peak chromatograms, selective transamination converted each peptide to a mixture of reaction products. In principle, selective transamination leads

to a net mass shift of -1.03 Da due to the removal of both the α-amino group and a hydrogen atom and then the addition of an oxygen atom to the α-carbon (Figure 4.27A). For both peptides, the correct transamination product was detected with the expected mass shift of -1.03 Da. However, the +74 Da by-products were also identified after the transamination of both peptides. As described by Sonomura *et al.* (2009b), formation of such by-products "can be explained by the addition of glyoxylate (before dehydration to the Schiff base)" where glyoxylate is the amino acceptor and the Schiff base is suggested as a +56 Da by-product (Figure 4.27B). Another side product of rat renin substrate was also identified with a mass shift of -45 Da, which may be explained by the decarboxylation of the N-terminal D residue (Figure 4.27C; Gilmore *et al.*, 2006). It should be noted that the decarboxylated peptide has also undergone transamination (mass shift = -1.03 Da), which explains the net mass shift of -45 Da instead of -44 Da (due to decarboxylation alone).

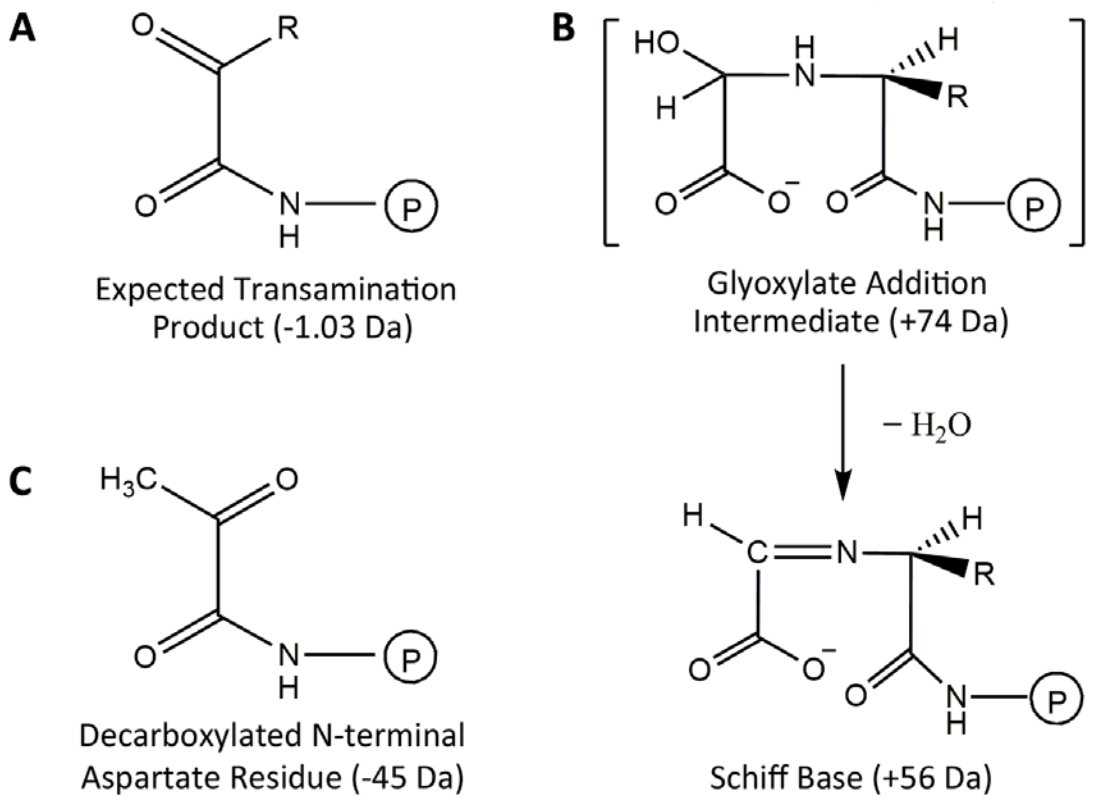


Figure 4.27 Chemical structures of the products of selective transamination on proteins/peptides. (**A**) The expected product of selective transamination with a mass shift = -1.03 Da. (**B**) The by-products resulting from glyoxylate addition (mass shift = +74 Da) and then Schiff base formation (mass shift = +56 Da). (**C**) The decarboxylation side product (mass shift = -45 Da, for N-terminal aspartate residue only). Ⓟ: a protein or peptide.

Prior to biotin tagging, another carbonyl-reactive compound, namely $BnONH_2$, was first employed to survey the chemical reactivity of the transaminated peptides. In previous studies on selective transamination, treatments with alkoxyamine compounds such as $BnONH_2$ were routinely performed to block the reactive carbonyl groups as they readily dimerised through aldol reactions (Papanikos *et al.*, 2001). After selective transamination and treatment with $BnONH_2$, both human ACTH and rat renin substrate exhibited the expected mass shift (+104.03 Da) when compared to their native counterparts. As revealed by XIC, both the $BnONH_2$-modified human ACTH and rat renin substrate were detected as a pair of peaks with slightly different signal intensities, which was likely due to the stereochemistry of the reaction products. The two stereoisomers might be produced in unequal quantities and separable by LC (Bachmann and Barton, 1938, Matlin *et al.*, 1990). Consistent with the report by Gilmore *et al.* (2006), $BnONH_2$ was also attached to the decarboxylated rat renin substrate since it also possessed a reactive carbonyl group (Figure 4.27C). In contrast, the +74 Da by-product was not modified with $BnONH_2$, confirming that this by-product did not contain any reactive carbonyl groups.

The newly imposed carbonyl group at the N-terminus of a transaminated peptide was then exploited for biotin tagging. As discussed above, an alkoxyamine-biotin tag was employed in the peptide test and was demonstrated to afford specific tagging of the transaminated N-terminus of human ACTH. In contrast, the biotinylated rat renin substrate could not be identified by the Mascot search. The biotinylated peptide was suggested by the error-tolerant search to exhibit additional mass shifts (e.g. -44 Da) due to decarboxylation or other unsuspected modifications. The large extent of decarboxylation may account for the failure to identify the expected biotinylation product of rat renin substrate. In addition, from a pragmatic viewpoint it is clear that transamination did not proceed with high efficiency on all substrates. Therefore, the yield of transamination may be another limiting factor in the identification of such peptides.

The difficulties in identifying the expected biotinylated peptides may also lie in precursor ion fragmentation. The tandem mass spectrum of the biotinylated human ACTH showed an unidentified fragment ion (m/z = 270.13, z = 1) with high intensity in addition to the peptide backbone fragments ($y_3 – y_8$). It was proposed that this fragment ion signified the cleavage of the attached biotin tag itself, which contains a linker region known to easily fragment (Figure 4.28; Tutturen *et al.*, 2013). Cleavage of the biotin tag did not interfere with the identification of the biotinylated human ACTH as it solely relied on the assignment of *y*-series ions, which in theory should not contain the biotin tag. In contrast, the biotin tag cleavage would

significantly impact the identification of the biotinylated rat renin substrate since *b*-series ions would be predominant in this case. The fragment ion at *m*/*z* = 270.13 was exploited by Tutturen *et al.* (2014) as a marker ion to aid in confident peptide identification. Similarly, a recent proteomic study on protein biotinylation used a different set of marker ions derived from biotin tag cleavage to identify biotinylated peptides (Udeshi *et al.*, 2017).

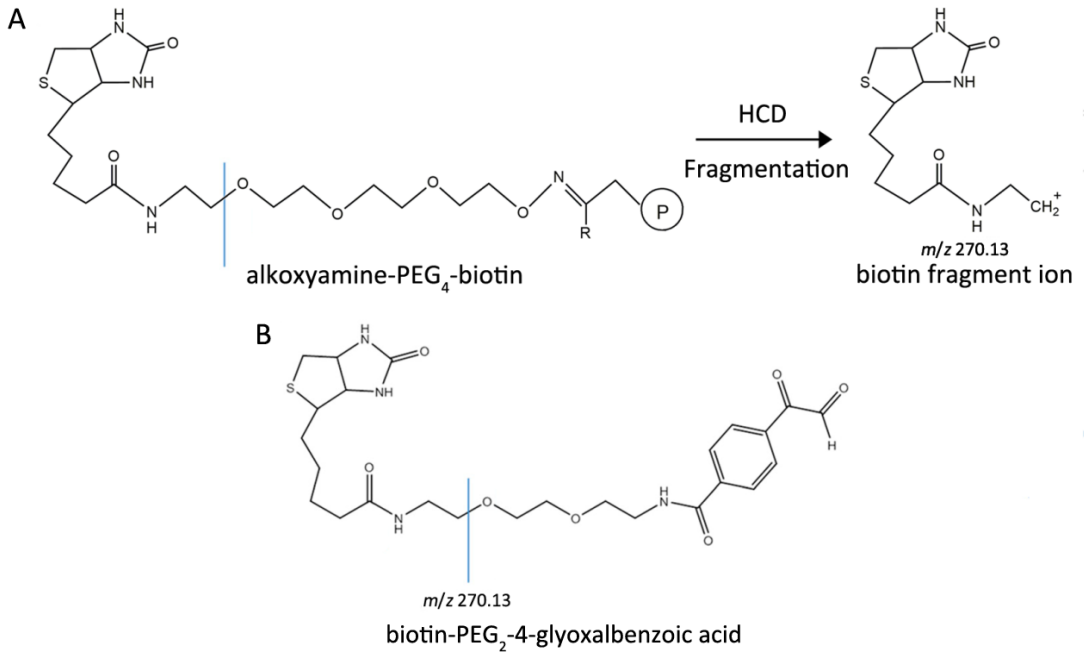


Figure 4.28 (**A**) Proposed cleavage of an alkoxyamine-biotin tag (e.g. alkoxyamine-PEG_4-biotin) attached to the N-terminus of a transaminated protein/peptide. The *m*/*z* value of the putative cleavage product is 270.13 (*z* = 1). (**B**) Similar fragmentation of biotin-PEG_2-4-glyoxalbenzoic acid (a citrulline-reactive biotin tag), as proposed by Tutturen *et al.* (2014). The cleavage sites are indicated by the blue lines. Ⓟ: a protein or peptide; *m*/*z*: mass-to-charge ratio; PEG: polyethylene glycol.

The partial success of the peptide test was also translated to the subsequent test on two model proteins: lysozyme C and BSA. With the knowledge acquired from the peptide test, the $BnONH_2$-modified N-terminal peptide of each protein was identified by Mascot searches using a pre-defined mass shift of +104.03 Da. In addition, an error-tolerant search suggested that the decarboxylated N-terminal peptide of BSA was also amenable to $BnONH_2$ modification. As explained above, the carbonyl groups introduced by transamination immediately require further blocking due to their susceptibility to aldol-type dimerisation. It is less of a problem for the transaminated peptides, which can be analysed in a timely fashion. However, the transaminated proteins were subjected to prolonged incubation (i.e. protein denaturation and proteolysis) before LC-MS/MS. Dimerisation of the transaminated N-termini might thus become significant. Blocking with $BnONH_2$ helped to resolve this issue, despite at the risk of a greater sample loss due to a second reaction and purification step.

In spite of the success with $BnONH_2$ modification of the transaminated protein N-termini, initial attempts to identify the protein N-termini tagged with hydrazide-biotin were proven largely unsuccessful. The negative result from the LC-MS/MS analysis may be partially rationalised by the low yields of transamination and biotin tagging. Incomplete transamination has been discussed above. With respect to biotin tagging, hydrazide-biotin was shown to have the lowest tagging efficiency among the four biotin reagents tested by Coffey and Gronert (2016). In comparison, the tagging efficiency of hydrazide-biotin was only 25 % relative to that of the best performer, i.e. alkoxyamine-PEG_4-SS-PEG_4-biotin.

Interestingly, dot blotting showed an opposite result of tagging the transaminated protein N-termini with hydrazide-biotin. The discrepancy between the LC-MS/MS and dot blotting results may be explained in two ways. First, the biotinylated N-termini might have undergone unsuspected modifications that conferred additional mass shifts. For lysozyme C, an error-tolerant search indeed assigned putative biotinylated N-terminal peptides, which had also undergone carbamylation, carbamidomethylation, etc. In the test with model proteins, lysozyme C and BSA were routinely denatured in urea solution to facilitate selective transamination. It is well acknowledged that urea solution can cause carbamylation of K, arginine (R), and protein N-termini, especially at elevated temperatures (Sun *et al.*, 2014). Meanwhile, excess iodoacetamide likely caused carbamidomethylation of K, methionine (M), or histidine (H) residue (Boja and Fales, 2001). Therefore, such assignments may be genuine, and caution should be exercised when using urea or iodoacetamide in future experiments.

Second, the hydrazone bond formed between the carbonyl group and hydrazide-biotin might have dissociated before LC-MS/MS analysis. Shi *et al.* (2015) reported that hydrazone bonds were labile to hydrolysis in mild acidic environments (pH < 5). After protease digestion, the biotinylated N-terminal peptides were indeed exposed to formic acid (0.1 – 0.5 %, v/v) during sample preparation and LC-MS/MS. This acidic environment might have caused accidental release of the biotin tag so that the biotinylated protein N-termini could not be identified. In comparison, oxime bonds formed by the reaction between alkoxyamines and carbonyl groups are more resistant to acid-catalysed hydrolysis (Kalia and Raines, 2008). Through direct comparisons of the hydrolytic stability between oxime and hydrazone linkages, the authors concluded that oxime bonds were nearly 10^3-fold more stable than the equivalent hydrazone bonds. In view of this claim and that alkoxyamines outperformed their hydrazide counterpart in tagging protein carbonyl groups (Coffey and Gronert, 2016), hydrazide-biotin was replaced by alkoxyamine-PEG_4-biotin to tag the transaminated protein N-termini. Again, the success with biotin tagging was supported by the dot blotting results.

Following the confirmation of biotin tagging on intact proteins, protease digestion and LC-MS/MS were performed once again in order to identify the biotin-tagged protein N-termini. Mascot searches identified the biotinylated N-terminal peptide for both lysozyme C and BSA, demonstrating the feasibility of the selective transamination approach for tagging protein N-termini. However, it should be noted that Nt-biotinylation was very difficult to identify in the case of lysozyme C and the identification was less reliable than that for BSA. This issue may be related to the low quality of the corresponding tandem mass spectra as well as the efficiency of transamination on N-terminal K residues, the latter of which has been reported to be very low (Sonomura *et al.*, 2009b). With respect to BSA, the decarboxylated protein N-termini were also amenable to biotin tagging through oxime bond formation.

The success with protein Nt-biotinylation led to the final test: enrichment of the biotin-tagged N-terminal peptides by AP. This test was divided into two parts: the first part involved intact protein AP where the biotinylated proteins were directly purified by immobilised avidin resins and analysed by SDS-PAGE; the second part required protease digestion of the biotinylated proteins, and the resulting peptides were subjected to AP and LC-MS/MS. Initially, NeutrAvidin agarose was the preferred avidin resin for enriching the biotinylated BSA. SDS-PAGE revealed that the biotinylated BSA could be enriched by this avidin resin. However, it was considered to be unsuitable for the test on lysozyme C since the denaturing conditions to elute the bound proteins likely caused the dissociation of tetrameric NeutrAvidin into monomers, which could interfere with the analysis of lysozyme C by SDS-PAGE. As a substitute for NeutrAvidin agarose, monomeric avidin agarose was demonstrated by SDS-PAGE to provide effective enrichment of the biotinylated lysozyme C.

An interesting observation from the above results was the detection of a biotin signal/protein band in a negative control (biotinylation-only) by dot blotting and SDS-PAGE, respectively. Protein tagging and AP often suffer from nonspecific reactions (Dubrovska, 2009). In addition, endogenously biotinylated proteins present a further source of interference with the biotin-(strept)avidin system (Tytgat *et al.*, 2015). However, it was postulated that specificity was not the primary issue in this case. Neither was natural protein biotinylation, since lysozyme C and BSA are not endogenously biotinylated according to the Swiss-Prot annotations. Instead, the presence of the biotin signal/protein band may be attributed to spontaneous protein carbonylation.

Carbonylation is a major form of oxidative modification caused by both physiological (e.g. redox signalling) and pathological (e.g. oxidative damage) processes (Wehr and Levine, 2012). Previous studies have shown that K, R, proline (P), threonine (T), and tryptophan (W) residues are the primary sites of *in vivo* protein carbonylation (Cabiscol *et al.*, 2014). However, the chemical structures of such carbonylated residues can be highly complex as a result of the multiple routes for protein carbonylation: I. peptide backbone cleavage; II. side chain oxidation; III. attachment of carbonyl-containing lipids; IV. protein glycoxidation, i.e. the combination of glycation and further oxidation (Fedorova *et al.*, 2014). For simplicity, only the products of side chain oxidation are shown in Figure 4.29.

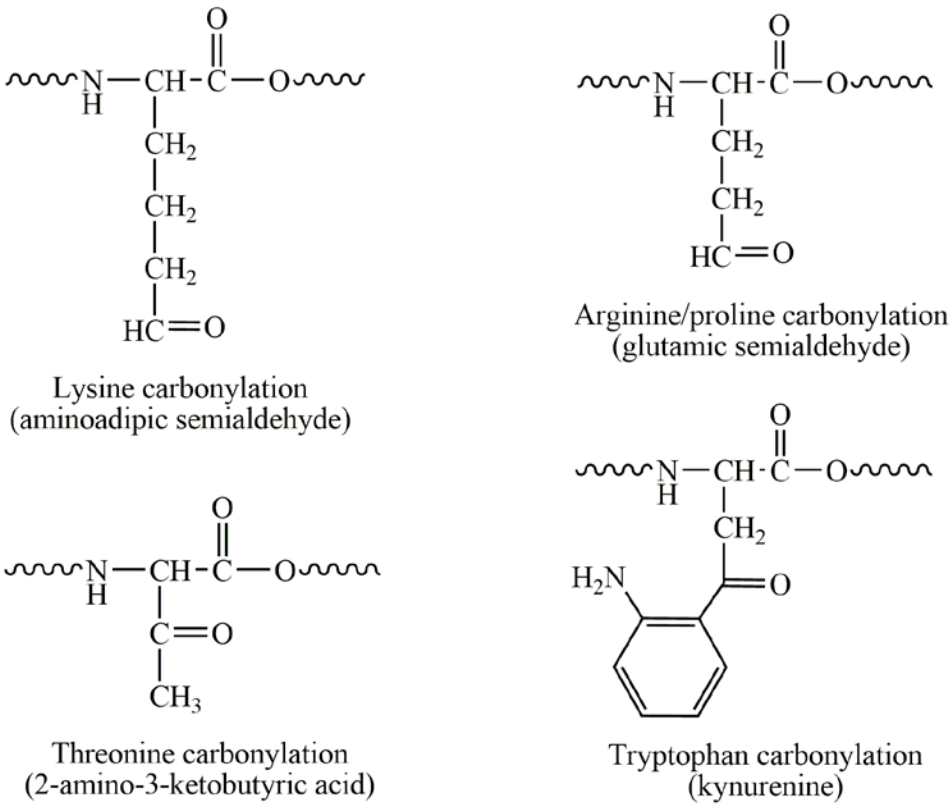


Figure 4.29 Chemical structures of carbonylated lysine (K), arginine (R), proline (P), threonine (T), and tryptophan (W) residues. The names of the carbonylation products are indicated in parentheses.

In the present study, however, the most likely cause of spontaneous protein carbonylation is the long-term storage of the purified proteins in the presence of air. Background carbonylation was indeed detected in other commercial proteins, supporting the above argument (Coffey and Gronert, 2016). Nevertheless, spontaneous protein carbonylation may become less of an issue when proteins are prepared freshly, especially in the presence of reducing agents and metal ion chelators (Rogowska-Wrzesinska *et al.*, 2014).

The second part of the final test was to identify the biotinylated N-terminal peptides after positive selection by AP. However, the success with the protein AP could not be reproduced at the peptide level. For BSA, only internal tryptic peptides were identified by the Mascot search after AP. Meanwhile, internal peptides and the native N-terminal peptide were identified in the lysozyme C samples after AP. The detection of internal peptides after AP may reflect an inadequate washing step, undesired specificity of this avidin resin, or simply the higher "flyability" of such peptides compared to the biotinylated N-terminal peptides. Nonetheless, an error-tolerant search identified the biotinylation product of the decarboxylated N-terminal peptide of BSA.

In conclusion, the present study employed a systematic test at both the peptide and protein levels to evaluate the feasibility of selective transamination for tagging and enriching the N-termini of proteins. One of the model peptides, human ACTH (with an N-terminal S residue), was readily transaminated and tagged with alkoxyamine-based biotin at the N-terminus, despite the presence of reaction by-products (e.g. mass shift = +74 Da). This result is in agreement with the previous findings that the transamination of N-terminal S residues was efficient (Sonomura *et al.*, 2009b). On the other hand, it proved to be difficult to obtain the same result on rat renin substrate and the N-terminal peptide of BSA, both starting with a D residue. This is likely due to the decarboxylation of such peptides to a large extent. Using error-tolerant searches, the decarboxylated peptides were suggested to react with the alkoxyamine-based biotin tag. In addition, the N-terminal peptide of lysozyme C (with an N-terminal K residue) was also amenable to transamination and biotin tagging, but the detection of this peptide in its native state even after AP suggests that the transamination of this peptide was incomplete. Indeed the N-terminal K residue was suggested to be one of the residues that were difficult to transaminate (Sonomura *et al.*, 2009b).

To our knowledge, this study may represent the development of the first N-terminalomic technique that employs carbonyl-reactive chemistry. However, the present study also revealed major issues in the selective transamination approach, namely the yields of transamination and biotin tagging. Due to time constraints, these issues were not thoroughly investigated in a systematic manner. Therefore, future studies should focus on improving the reaction yields and minimising sample loss during protein purification. Experiment parameters to be refined include: I. the concentration of the protein substrate, the amino acceptor, the metal ion catalyst, and the base; II. reaction duration and temperature; III. addition of a metal ion chelator after transamination; IV. use of alternative protein purification techniques.

In addition, the biotin tagging and AP need to be optimised as well. Since the pH of biotin tagging (6.5) was empirically determined, the refinement of this parameter should be emphasised. In the present study, protein and peptide AP produced mixed results: although effective AP of the biotinylated proteins was achieved, the same procedure on proteolytic peptides only showed high background binding but not the enrichment of the biotinylated N-terminal peptides. This procedure should benefit from the use of immobilised avidin beads with minimal nonspecific binding (i.e. NeutrAvidin agarose) and a biotin tag that can be conveniently eluted, including cleavable biotin tags (other than the thiol-cleavable ones) and modified biotin tags with a reduced affinity (e.g. desthiobiotin). This combination is also preferred by other positive selection strategies including N-CLAP, *O*-methylisourea, and Subtiligase (Xu and Jaffrey, 2010, Timmer and Salvesen, 2011, Wiita *et al.*, 2014).

Since selective transamination and biotin tagging produced heterogeneous products, the current spectral acquisition method and data analysis settings should be modified to account for additional mass shifts. Specifically, identification of the biotin-tagged peptides should benefit from targeted proteomics that screens for the potential marker ion at m/z = 270.13. Since the marker ion likely results from the cleavage of an attached biotin tag, any precursor ion that gives rise to this fragment ion (with high intensity) may correspond to a biotin-tagged peptide. For instance, Tutturen *et al.* (2014) took advantage of this marker ion to show that the combination of a citrulline-reactive biotin tag and streptavidin beads provided effective enrichment of citrullinated peptides: ≥ 95 % of the resulting tandem mass spectra contained the marker ion at m/z = 270.13, which was only present in 1 % of the tandem mass spectra in the control group (i.e. without enrichment). The importance of diagnostic marker ions has also been highlighted by the recent study that employed biotin-specific antibodies to enrich and globally identify biotinylated peptides (Udeshi *et al.*, 2017).

In view of the difficulties to enrich for the biotinylated protein N-termini under the current conditions, the original scheme of the selective transamination approach was modified to exclude the AP step for subsequent experiments. Therefore, the modified approach only included four steps: transamination, biotinylation, protease digestion, and LC-MS/MS analysis. As described in Chapter 5, the modified approach was applied on a highly complex mixture of proteins (extracted from Jurkat T-lymphocytes) in order to globally identify protein N-termini. In addition, the feasibility of selective transamination for improving the proteome coverage by shotgun proteomics was also evaluated.

Chapter 5. Use of the transamination approach in shotgun proteomics

5.1 Introduction

Jurkat T-lymphocytes are a cancer cell line established from a patient with acute T-cell leukaemia (Schneider *et al.*, 1977). With an exceptional ability to synthesise interleukin-2 (IL-2), Jurkat T-cells have been used primarily as a model system to dissect the molecular mechanism associated with T-cell receptor (TCR) signalling pathways (reviewed in Weiss and Stobo, 2015). For instance, Jurkat T-cells played a pivotal role in identifying the two-stimulus requirement for human T-cell activation (Weiss and Stobo, 1984). A variety of mutant cell lines have also been derived from wild-type Jurkat T-cells to further study the signalling cascades upon T-cell activation (Abraham and Weiss, 2004). To date, more than 17,000 references for "Jurkat" are documented in PubMed.

As described in Chapter 4, a "positive selection" N-terminalomic strategy was devised on the basis of selective transamination. This approach was then systematically investigated using model peptides and proteins. As demonstrated by these tests, the transamination approach can be employed to specifically attach an affinity tag to the amino (N)-terminus of a peptide or protein. In principle, the N-terminal tag can facilitate positive selection and unambiguous identification of true protein N-termini through mass spectrometry (MS)-based proteomics. In addition to model peptides and proteins, the development of this approach requires further tests on a biological system that serves as a more complex model. The results of such tests will shed light on the feasibility of this approach for proteome-wide analysis.

Jurkat T-cells are an ideal source of complex proteomes as they have been experimentally verified to express > 8,000 proteins (Geiger *et al.*, 2012). In addition, these cells are robust and simple to handle in laboratory, and their expansion rate is high (population doubling time = 48 hours). Compared with the classical biochemical studies mentioned above, the proteomic analysis of this cell line is a more recent topic well into the post-genomic era (Kang *et al.*, 2005). Given the importance of (de)phosphorylation events in cell signalling, Jurkat proteomics mainly focuses on the identification of proteins affected by this post-translational modification (PTM) and the changes in their phosphorylation states (Nguyen *et al.*, 2016). In contrast, a comprehensive analysis of the Jurkat T-cell proteome seems to have been given less attention.

With respect to N-terminalomics, Jurkat T-cells have served as a robust model system of apoptosis to demonstrate the feasibility of several positive and negative selection approaches for global mapping of proteolytic processing events (described in section 1.5). These N-terminalomic approaches include N-terminal COFRADIC (combined fractional diagonal chromatography), N-CLAP (N-terminalomics by chemical labeling of the α-amine of proteins), and Subtiligase (reviewed in Plasman *et al.*, 2013). In addition to proteolytic processing, the extent of N-terminal methionine excision (NME) in the Jurkat proteome was also surveyed using the N-CLAP approach (Xu *et al.*, 2009). These studies demonstrated the utility of Jurkat T-cells in the development of N-terminalomic approaches.

For the above reasons, Jurkat T-cells were also employed for testing the transamination approach. The present study sought to combine this approach with liquid chromatography–tandem mass spectrometry (LC-MS/MS) in order to identify the N-termini of proteins expressed in Jurkat T-cells. This study was carried out with a clear emphasis on the identification of "free" protein N-termini, which refer to those without any PTMs. Similar to the previous tests with model proteins, Jurkat proteins were also transaminated under the "salt-free" condition (Sonomura *et al.*, 2009a), but urea was replaced with guanidine hydrochloride (HCl) to avoid protein carbamylation (see sections 2.8 & 4.3).

Additionally, it is well known that undersampling is currently a major problem in shotgun proteomics (Wang *et al.*, 2010). This is partly because of the wide range of protein concentration in certain proteomes. In the shotgun analysis of such proteomes, peptide ions derived from highly abundant proteins dominate over the low-abundance ones when conventional data-dependent acquisition (DDA) is adopted. In certain types of protein samples, e.g. unfractionated blood plasma, up to 50 % of the MS machine time is spent analysing just 22 proteins, which account for 99 % of the blood proteome by mass (Wildes and Wells, 2010). A related issue is that certain peptides are not detected by LC-MS/MS, e.g. due to their "stickiness" which may cause them to adhere to surfaces prior to MS (Maes *et al.*, 2014), or because they ionise poorly (Mirzaei and Regnier, 2006).

Therefore, a second aim of the present study was to test the idea that transamination on a mixture of proteolytic peptides may help overcome the undersampling problem and thus increase proteome sequence coverage. Specifically, the two hypotheses are: I. with peptide transamination, shotgun proteomics will identify a different but overlapping set of Jurkat proteins relative to those identified without this treatment; and II. combining these two datasets will expand the surveyed Jurkat proteome. Peptide transamination also followed the protocol by Sonomura *et al.* (2009a) to maintain consistency throughout this study.

5.2 Results

5.2.1 Experiments with complex protein mixtures from Jurkat T-cells

Experiments using model peptides and proteins were detailed in Chapter 4. The present chapter will describe an investigation of selective transamination on a complex mixture of proteins that were extracted from Jurkat T-cells. Jurkat cell cultures were maintained in our lab and their morphology was routinely verified by microscopy (Figure 5.1).

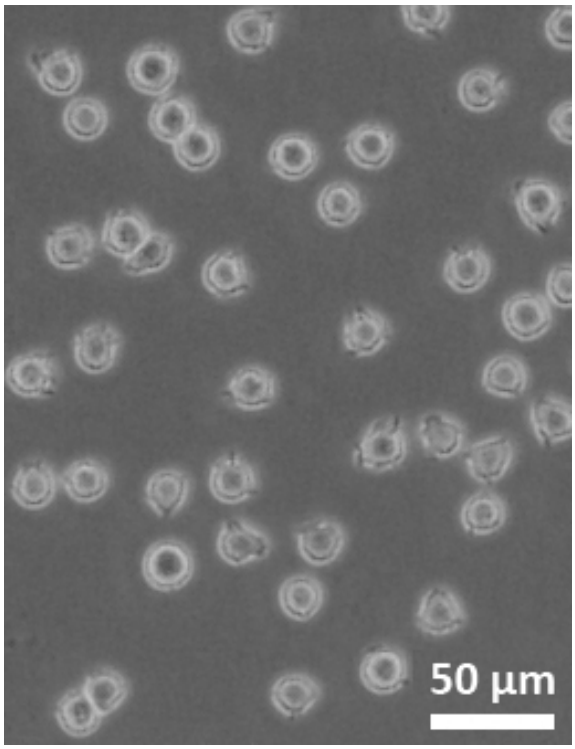


Figure 5.1 Representative microscopic image of Jurkat T-cells at low density (1 x 10^5 cells/ml). Scale bar = 50 µm.

Initially, crude protein extracts from Jurkat T-cells were subjected to selective transamination immediately after the estimation of protein concentrations using a bicinchoninic acid (BCA) assay. After transamination, the crude protein samples were allowed to react with 2,4-dinitrophenylhydrazine (DNPH), which is routinely used in determining protein carbonyl content (Wehr and Levine, 2012). The carbonyl content in the treated sample (Trans+, DNPH+) and three negative controls (native protein, DNPH-only, and transamination-only) was quantified by a spectrophotometric assay that measures the absorbance at 370 nm (A_{370}). However, the result of the DNPH assay showed uniform A_{370} readings across all four samples (data not shown).

Through a literature search, genomic DNA fragments that remained in the crude Jurkat protein extracts were identified as a potential source of interference in the DNPH assay. Previous studies have established that DNA is carbonyl-positive and thus readily reacts with

DNPH (Luo and Wehr, 2009). Therefore, contamination with sheared genomic DNA likely rendered the DNPH assay an ineffective measurement of carbonyl content in the crude Jurkat protein extracts. More importantly, the remaining DNA fragments also had the potential to interfere with downstream modifications (e.g. biotin tagging) of the carbonyl groups introduced by transamination.

In view of the risks posed by these DNA fragments, the crude protein extracts were treated with DNase I and then subjected to gel filtration (7K MWCO) to remove genomic DNA. Samples of the Jurkat protein extracts with or without DNA removal were compared by agarose gel electrophoresis. As shown in Figure 5.2, genomic DNA was readily detected in the crude protein extracts as smearing bands on the gel before treatment with DNase I. After the treatment, such bands were nearly absent in the protein sample. Based on this result, it was concluded that the DNase I-treated protein samples were free from genomic DNA interference and could be used for further analyses.

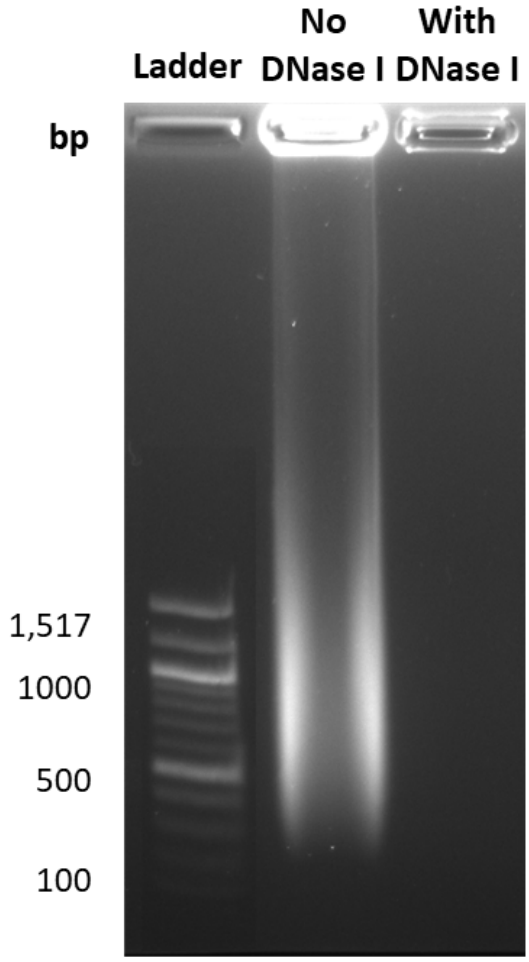


Figure 5.2 Gel electrophoresis of genomic DNA in crude Jurkat protein extracts before and after DNA removal with DNase I. Equal amounts of the protein samples (based on protein concentration estimation) were separated on a 1 % (w/v) agarose gel, and then stained with GelRed. bp: base-pair.

The DNA-free protein samples were again transaminated and treated with DNPH, but the A_{370} assay for protein carbonylation quantitation was replaced by dot blotting, which used an antibody specific for dinitrophenol groups. As shown in Figure 5.3, the dot blotting results showed that the dinitrophenol signal was detected in the treated sample (Trans+, DNPH+) with high intensity. In contrast, three negative controls, i.e. native protein (Trans-, DNPH-), DNPH-only (Trans-, DNPH+), and transamination-only (Trans+, DNPH-) merely exhibited weak dinitrophenol signals. These results were largely consistent with those from the previous tests with chicken egg-white lysozyme (lysozyme C) and bovine serum albumin (BSA), which were also DNA-free (see Figures 4.17 & 4.19). In conclusion, selective transamination was also capable of introducing carbonyl groups to a complex mixture of proteins extracted from a cell line, and DNA removal was required for eliminating any interference with the detection of the introduced carbonyl groups.

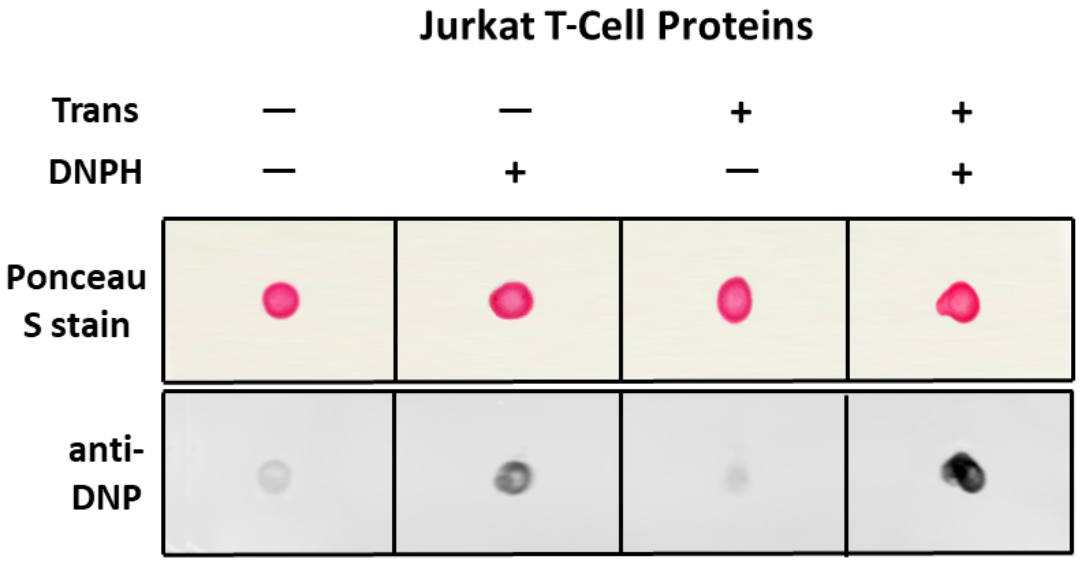


Figure 5.3 Dot blotting of the transaminated and DNPH-treated proteins that were extracted from Jurkat T-cells. In addition to the treated sample (Trans+, DNPH+), three analysed negative controls are: native protein (Trans-, DNPH-), DNPH-only (Trans-, DNPH+), and transamination-only (Trans+, DNPH-). Trans: transamination; DNPH: 2,4-dinitrophenolhydrazine; DNP: dinitrophenol groups.

Having established that selective transamination could be employed to introduce carbonyl groups to Jurkat protein extracts, the DNase I-treated protein samples were transaminated and the resulting carbonyl groups at protein N-termini were tagged using a carbonyl-specific biotin reagent (Figure 5.4). The biotin-tagged proteins were then analysed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and Western blotting.

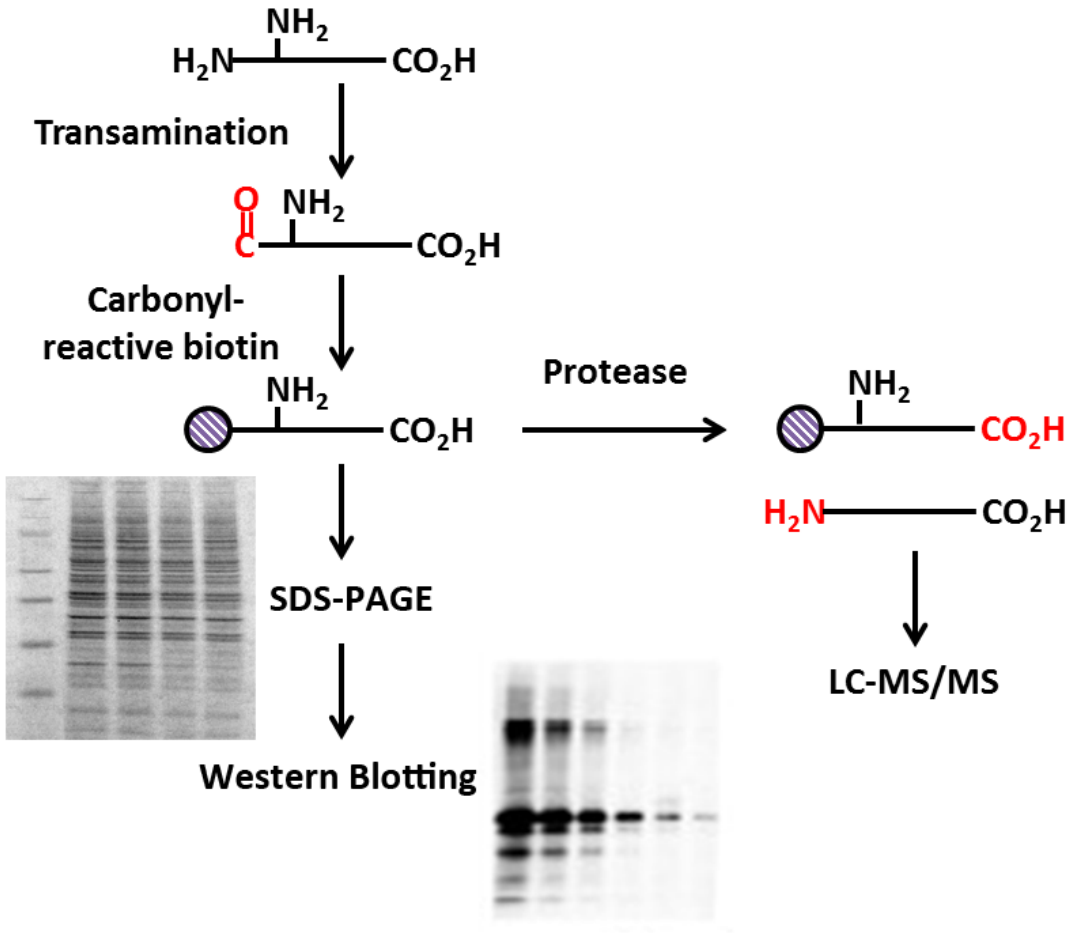


Figure 5.4 Strategy used for selective transamination of proteins in complex mixtures. The workflow consists of five major steps: I. selective transamination; II. carbonyl-specific biotinylation; III. SDS-PAGE and Western blotting; IV. protease digestion; V. LC-MS/MS analysis. The red labels indicate new chemical groups introduced by selective transamination or protease digestion, whereas the hatched circles indicate the N-terminal biotin tags introduced through carbonyl-reactive chemistry (i.e. oxime bond formation). SDS-PAGE: sodium dodecyl sulfate–polyacrylamide gel electrophoresis; LC-MS/MS: liquid chromatography–tandem mass spectrometry.

In each case, after transamination and tagging with alkoxyamine-PEG_4-biotin, a BCA assay was carried out on the protein samples to estimate protein concentration prior to SDS-PAGE. As shown in Figure 5.5 (left), four protein samples were analysed: native protein (Trans-, Biotin-), biotinylation-only (Trans-, Biotin+), transamination-only (Trans+, Biotin-), and the treated sample (Trans+, Biotin+). Following SDS-PAGE, the same set of protein samples were subjected to Western blotting that employed an antibody specific for the biotin tag. The objective was to evaluate the efficiency and selectivity of the transamination approach on complex protein mixtures. Given that all the Jurkat samples were at the same protein concentration, the differences (if any) in the intensity of biotin signals should reflect the efficacy of transamination and biotin tagging.

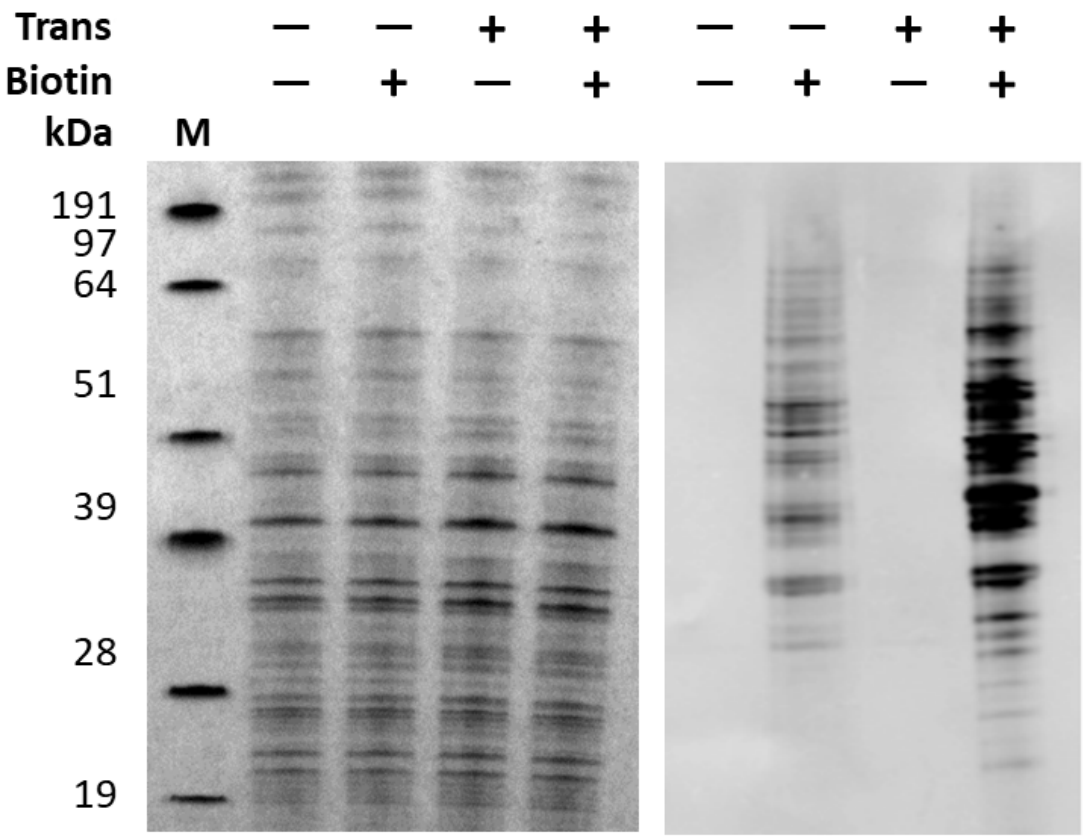


Figure 5.5 SDS-PAGE (left) and Western blotting (right) of the transaminated and biotin-tagged proteins from Jurkat T-cells. The carbonyl-specific biotin reagent used was alkoxyamine-PEG_4-biotin. Equal amounts of four protein samples were separated on a 10 % NuPAGE™ Bis-Tris gel. The treated sample (Trans+, Biotin+) was analysed together with three negative controls: native protein (Trans-, Biotin-), biotinylation-only (Trans-, Biotin+), and transamination-only (Trans+, Biotin-). SDS-PAGE: sodium dodecyl sulfate–polyacrylamide gel electrophoresis; Trans: transamination; M: protein molecular weight marker.

As expected, the biotin signal was absent in the two protein samples without biotin tagging, i.e. native protein (Trans-, Biotin-) and transamination-only (Trans+, Biotin-). In contrast, the signal was present in one of the negative controls, biotinylation-only (Trans-, Biotin+). However, the biotin signal was detected in the treated sample (Trans+, Biotin+) with much higher intensity (Figure 5.5, right). These results were consistent with those obtained via dot blotting of the biotin-tagged model proteins (see Figures 4.17 & 4.19). Therefore, selective transamination was confirmed to enable the addition of carbonyl-specific biotin tags to the N-termini of proteins in a complex mixture (e.g. Jurkat whole-cell extracts).

Having tested the transamination approach on Jurkat protein extracts through SDS-PAGE and Western blotting analyses, this approach was then applied to study two aspects of the Jurkat T-cell proteome: identification of free or acetylated protein N-termini, and comprehensive profiling of the Jurkat proteome by shotgun proteomics.

5.2.2 Proteomic identification of the N-termini of Jurkat proteins using selective transamination

MS-based proteomic analyses are routinely performed for various cell and tissue types in our group. Without prior fractionation or derivatisation, a collection of > 1,000 proteins can be robustly identified using the shotgun approach. MS-based proteomics is particularly well suited to detecting different types of protein PTMs on top of protein identification. In terms of T-cells, proteomic studies have mainly focused on the identification of protein phosphorylation, since this reversible PTM plays a key role in the TCR signalling switched on by T-cell activation (Huse, 2009).

On the other hand, PTMs at protein N-termini have not been well studied in Jurkat T-cells. Previous studies mainly employed N-terminalomic strategies to profile NME in the Jurkat proteome and to identify *neo*-N-termini that are indicative of proteolytic processing events (Van Damme *et al.*, 2005, Xu *et al.*, 2009). In contrast, proteins with free N-termini are relatively neglected since they only account for < 20 % of all human proteins (Giglione *et al.*, 2015). In principle, selective transamination is a promising strategy to identify free protein N-termini when combined with carbonyl-specific biotin tagging. The irreversibly attached biotin tag could aid in the unambiguous identification of free protein N-termini.

In the present study, a transamination experiment was carried out in which Jurkat proteins were subjected to selective transamination and then tagged with alkoxyamine-PEG_4-biotin, alkoxyamine-PEG_4-SS-PEG_4-biotin, or *O*-benzylhydroxylamine ($BnONH_2$) as described before. Subsequently, the biotin-tagged/$BnONH_2$-modified proteins were digested with trypsin for LC-MS/MS analysis. The resulting LC-MS/MS data (in the Orbitrap RAW format) were processed and searched against a local Swiss-Prot database (taxonomy = *Homo sapiens*) to detect biotin tags at the N-termini of Jurkat proteins. N-terminal (Nt)-acetylation was also included in the list of variable modifications. In contrast, the transamination step was replaced by incubation in ddH_2O in the control experiment where a database search was also performed to identify Nt-acetylated or free protein N-termini. In both experiments, the protein N-termini identified by Mascot were classified according to N-terminal PTMs, which will be described in detail later.

In the control experiment, overall 20,265 tryptic peptides were reliably detected (peptide expectation value, *E*-value ≤ 0.05) from 21 biological replicates (N = 21). Figure 5.6 shows the cumulative addition of unique tryptic peptides by each replicate. It suggests that peptide detection approached saturation after the 19^{th} replicate analysis. As described previously, it

was difficult to perform a pair-wise comparison of the detected peptides for all 21 replicates. Nevertheless, a 6-set Venn diagram was drawn to compare the detected peptides between every two replicates out of six, which accounted for > 75 % of the total peptide number (see Appendix 1). Finally, a total of 1,880 Jurkat proteins were inferred from the detected peptides under the "two-peptide rule" (i.e. a protein must be identified on the basis of ≥ 2 significant peptide hits).

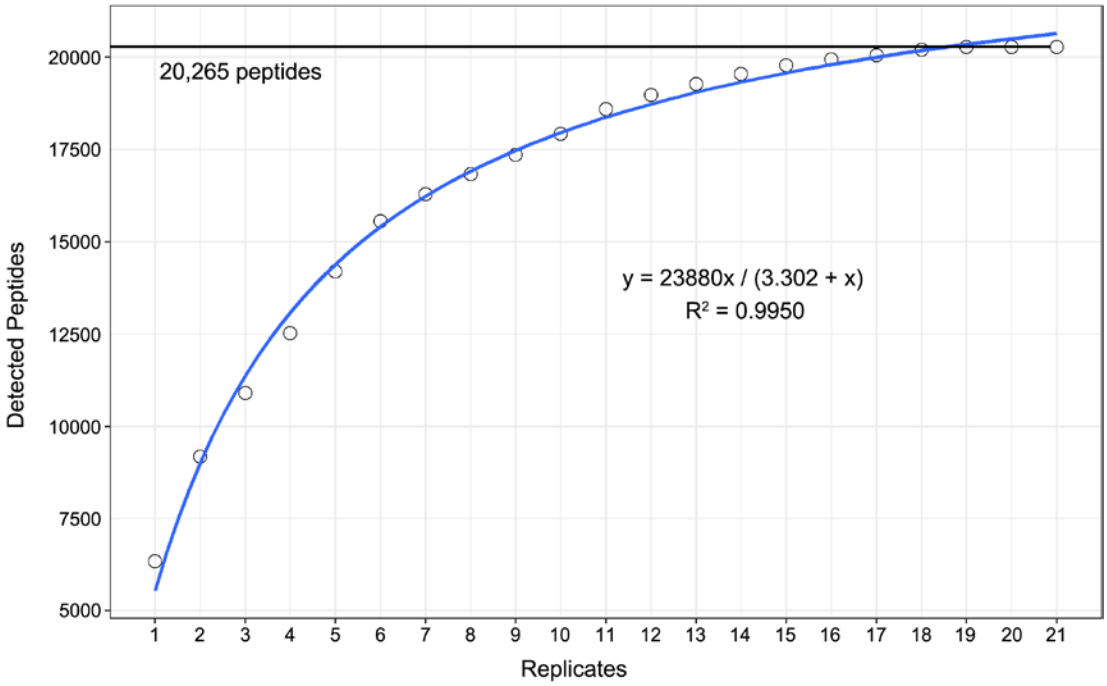


Figure 5.6 Saturation curve of unique tryptic peptides detected by replicate analyses in the control experiment. The curve (blue) was fitted to the experimental data through a nonlinear regression analysis in R; the total peptide number is indicated by the straight line in black.

Among the detected peptides, there were 568 N-terminal peptides assigned to 424 different proteins. As the local Swiss-Prot database was not featured with protein N-terminal annotations, the "N-terminal peptides" were defined here as the peptides starting with the first amino acid residue (always methionine, M) or the second one in their canonical sequences. Examples of the assigned N-terminal peptides are given in Table 5.1, and their respective protein sequences are shown in Figure 5.7. As mentioned above, these protein N-termini were further divided into two groups according to the state of N-terminal PTMs: "free" or "Nt-acetyl". There were 104 free protein N-termini (from 83 Jurkat proteins) and 469 N-terminal peptides (from 346 proteins) in the Nt-acetyl group.

Table 5.1 Examples of N-terminal peptides assigned to Jurkat proteins by Mascot database searches[a]. The corresponding amino acid sequences are shown in Fig. 5.7.

Protein Name	Swiss-Prot ID	N-terminal Peptide (Start – End)	Score	Duplicate PSM No.
Tubulin beta chain	TBB5_HUMAN (P07437)	MREIVHIQAGQCGNQIGAK (1 – 19)	94.23	117
Actin, cytoplasmic 1	ACTB_HUMAN (P60709)	DDDIAALVVDNGSGMCK (2 – 18) + Acetyl (N-term)	96.1	64

[a] Orbitrap RAW data were processed and searched against the local decoy database (Swiss-Prot, taxonomy = *Homo sapiens*). The N-terminal peptides of tubulin beta chain (Swiss-Prot ID: TBB5_HUMAN/P07437) and actin, cytoplasmic 1 (ACTB_HUMAN/P60709) start with the first (methionine, M) and the second residues (aspartate, D), respectively. Duplicate PSM No.: number of duplicate peptide-spectrum matches (PSM); Acetyl (N-term): N-terminal acetylation.

Tubulin beta chain OS=Homo sapiens GN=TUBB PE=1 SV=2
Protein sequence coverage: 86%

```
  1 MREIVHIQAG QCGNQIGAKF WEVISDEHGI DPTGTYHGDS DLQLDRISVY
 51 YNEATGGKYV PRAILVDLEP GTMDSVRSGP FGQIFRPDNF VFGQSGAGNN
101 WAKGHYTEGA ELVDSVLDVV RKEAESCDCL QGFQLTHSLG GGTGSGMGTL
151 LISKIREEYP DRIMNTFSVV PSPKVSDTVV EPYNATLSVH QLVENTDETY
201 CIDNEALYDI CFRTLKLTTP TYGDLNHLVS ATMSGVTTCL RFPGQLNADL
251 RKLAVNMVPF PRLHFFMPGF APLTSRGSQQ YRALTVPELT QQVFDAKNMM
301 AACDPRHGRY LTVAAVFRGR MSMKEVDEQM LNVQNKNSSY FVEWIPNNVK
351 TAVCDIPPRG LKMAVTFIGN STAIQELFKR ISEQFTAMFR RKAFLHWYTG
401 EGMDEMEFTE AESNMNDLVS EYQQYQDATA EEEEDFGEEA EEEA
```

Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1
Protein sequence coverage: 99%

```
  1 MDDDIAALVV DNGSGMCKAG FAGDDAPRAV FPSIVGRPRH QGVMVGMGQK
 51 DSYVGDEAQS KRGILTLKYP IEHGIVTNWD DMEKIWHHTF YNELRVAPEE
101 HPVLLTEAPL NPKANREKMT QIMFETFNTP AMYVAIQAVL SLYASGRTTG
151 IVMDSGDGVT HTVPIYEGYA LPHAILRLDL AGRDLTDYLM KILTERGYSF
201 TTTAEREIVR DIKEKLCYVA LDFEQEMATA ASSSSLEKSY ELPDGQVITI
251 GNERFRCPEA LFQPSFLGME SCGIHETTFN SIMKCDVDIR KDLYANTVLS
301 GGTTMYPGIA DRMQKEITAL APSTMKIKII APPERKYSVW IGGSILASLS
351 TFQQMWISKQ EYDESGPSIV HRKCF
```

Figure 5.7 Amino acid sequences of tubulin beta chain (Swiss-Prot ID: TBB5_HUMAN/P07437) and actin, cytoplasmic 1 (ACTB_HUMAN/P60709). The sequences shown in red were matched to the experimental data by Mascot database searches.

Notably, the sum of free (83) and Nt-acetylated (346) proteins was larger than the total number (424) of Jurkat proteins with assigned N-termini ($N_{free} + N_{acetyl} > N_{total}$). It led to the speculation that the N-termini of some proteins existed in both free and Nt-acetylated forms. By comparing these two datasets, five overlapping protein hits were indeed identified. These proteins were thus excluded from the list of proteins with free N-termini. Since Nt-acetylation is currently regarded as an irreversible PTM (Drazic *et al.*, 2016), this result suggested incomplete Nt-acetylation upon the synthesis of nascent proteins.

The remaining 78 proteins in the list were further cross-referenced with their corresponding entries in the online Swiss-Prot database, which is featured with N-terminal annotations. The N-termini of nine proteins identified in this study were shown to contradict with existing database annotations (Table 5.2). As a result, it was concluded that 69 out of the 78 Jurkat proteins possessed free N-termini. Two particular Jurkat proteins, cytoplasmic serine-tRNA ligase (Swiss-Prot ID: SYSC_HUMAN/P49591) and peptidyl-prolyl cis-trans isomerase FKBP5 (FKBP5_HUMAN/Q13451), were shown to have undergone NME. These novel discoveries were not documented in the Swiss-Prot database.

Table 5.2 List of the experimentally identified free protein N-termini that conflict with existing N-terminal annotations in the online Swiss-Prot database[a].

Protein Swiss-Prot ID	N-terminal Peptide (Start – End)	N-terminal PTM
GSTP1_HUMAN (P09211)	MPPYTVVYFPVR (1 – 12)	NME
SRP14_HUMAN (P37108)	MVLLESEQFLTELTR (1 – 15)	NME
RNH2C_HUMAN (Q8TDP1)	MESGDEAAIERHR (1 – 13) + Oxidation	Nt-acetyl
LSM4_HUMAN (Q9Y4Z0)	MLPLSLLK (1 – 8)	Nt-acetyl
H2B1O_HUMAN (P23527)	PDPAKSAPAPK (2 – 12)	NME & Nt-acetyl
FKBP4_HUMAN (Q02790)	TAEEMKATESGAQSAPLPMEGVDISPK (2 – 28)	NME & Nt-acetyl
SAE1_HUMAN (Q9UBE0)	VEKEEAGGGISEEEAAQYDR (2 – 21)	NME & Nt-acetyl
DOPD_HUMAN (P30046)	PFLELDTNLPANRVPAGLEK (2 – 21)	NME & Nt-acetyl
EIF2A_HUMAN (Q9BY44)	APSTPLLTVR (2 – 11)	NME & Nt-acetyl

[a] Orbitrap RAW data were processed and searched against the local decoy database (Swiss-Prot, taxonomy = *Homo sapiens*). After cross-referencing with the online N-terminal annotations, these protein N-termini were removed from the free dataset. PTM: post-translational modification; NME: N-terminal methionine excision; Nt-acetyl: N-terminal acetylation.

In summary, the control group consisted of 415 (69 + 346) instead of 424 Jurkat proteins with assigned N-termini (Figure 5.8). They constituted > 20 % of the overall identified proteins. In detail, Nt-acetylated proteins (346) and free proteins (69) accounted for 18 % and < 4 % of all the identified proteins, respectively. These figures were in stark contrast to those reported elsewhere (Giglione *et al.*, 2015). However, this discrepancy drastically diminished when only focusing on the 415 proteins with assigned N-termini: 20 % of them were free of N-terminal PTMs, whereas the remaining 80 % were blocked by Nt-acetylation.

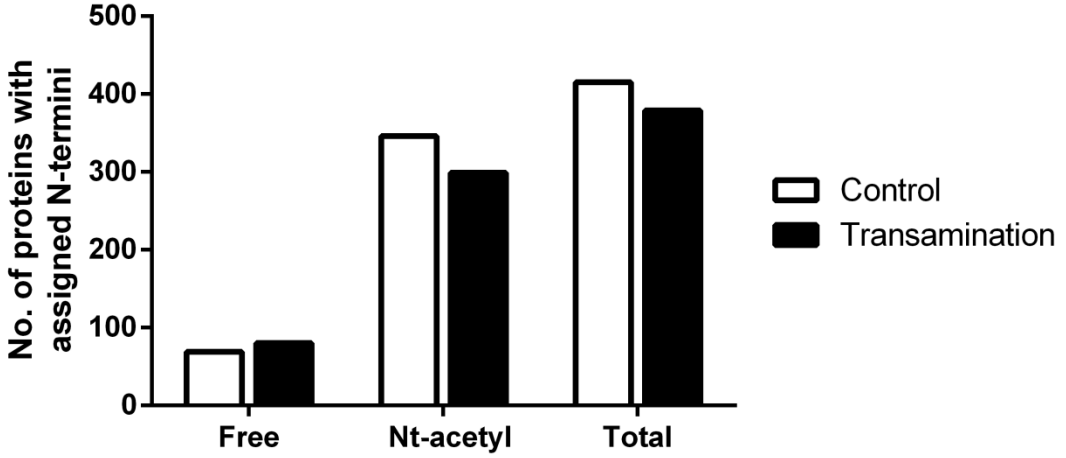


Figure 5.8 Number of Jurkat proteins with assigned N-termini in the control or transamination groups. The assigned protein N-termini are divided into "free" and "Nt-acetyl" groups based on their N-terminal post-translational modifications (PTMs). Nt-acetyl: N-terminal acetylation.

Proteomic analysis of the transaminated samples (N = 21) initially detected a total of 21,922 tryptic peptides with high confidence (*E*-value ≤ 0.05). Similar to the control group, peptide detection approached saturation by the 19th round of analysis of the transaminated samples (Figure 5.9). Furthermore, a Venn diagram was also drawn to compare the tryptic peptides detected by six replicates, which covered > 80 % of the total peptide number (see Appendix 1). The 21,922 peptides in turn led to the identification of 2,022 Jurkat proteins. Within this cohort, 489 N-terminal peptides were identified and assigned to 382 Jurkat proteins. These proteins were further divided into four categories according to the state of N-terminal PTMs: "free", "Nt-trans", "Nt-biotin", or "Nt-acetyl". "Nt-trans" referred to the proteins with transaminated N-termini or those further modified with $BnONH_2$, whereas the "Nt-biotin" group contained the proteins of which the N-termini were biotin tagged following transamination. Initially, there were 83 Jurkat proteins identified with free N-termini, 18 Nt-trans proteins, and 299 Nt-acetylated proteins. In terms of biotin tagging, only 4 out of 22 Nt-biotin proteins met the stringent criteria (*E*-value ≤ 0.05; the cognate proteins were identified with ≥ 2 significant peptides).

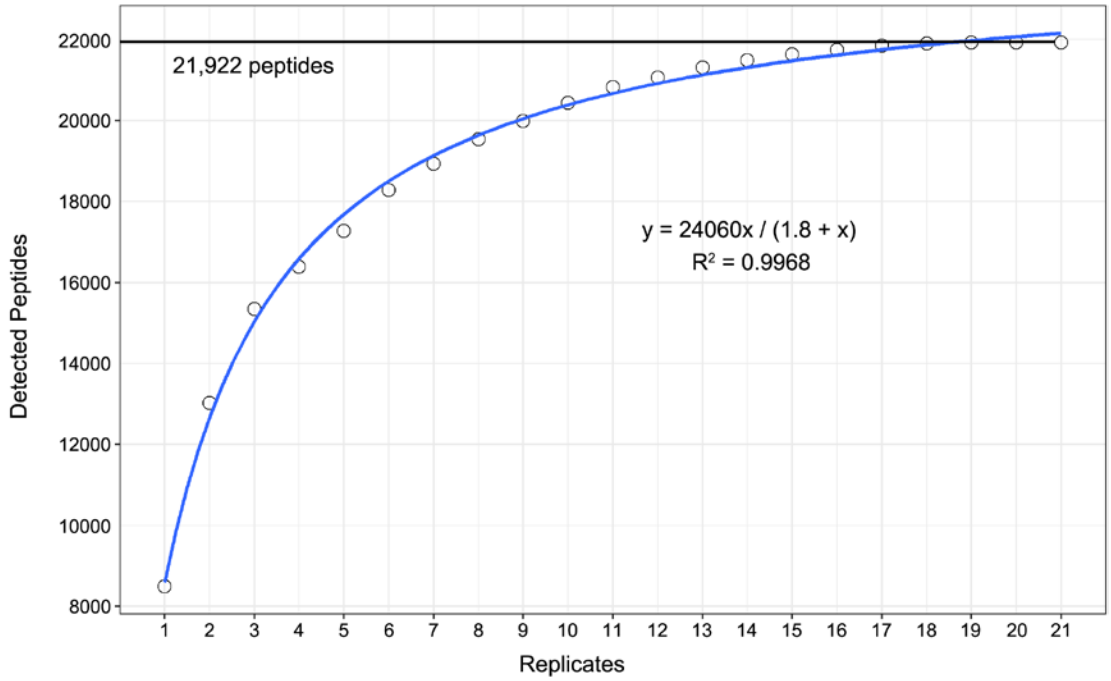


Figure 5.9 Saturation curve of unique tryptic peptides detected by replicate analyses in the transamination experiment. The curve (blue) was fitted to the experimental data through a nonlinear regression analysis in R; the total peptide number is indicated by the straight line in black.

Similar to the control group, the sum of proteins from all four categories was not equal to the total number of proteins with assigned N-termini. A hierarchical strategy was thus employed to generate a non-duplicate dataset of protein N-termini (Figure 5.10). First, the free, Nt-biotin, and Nt-trans proteins were merged to generate a combined free dataset. Overlapping protein hits were removed since in principle they were all derived from proteins with free N-termini. The combined free dataset was then compared with the Nt-acetyl dataset to remove protein hits that existed in both free and Nt-acetylated forms. Finally, cross-referencing with the online Swiss-Prot database was performed to remove protein hits that contradicted with existing N-terminal annotations.

Initially, the combined free dataset consisted of 88 Jurkat proteins. Five protein hits were then removed from this dataset as they existed in both free and Nt-acetylated forms. Cross-referencing with the online Swiss-Prot database resulted in further removal of three protein hits from the non-duplicate free dataset. These were ribonuclease H2 subunit C (RNH2C_HUMAN/ Q8TDP1), SUMO-activating enzyme subunit 1 (SAE1_HUMAN/ Q9UBE0), and D-dopachrome decarboxylase (DOPD_HUMAN/P30046). Consequently, there were 80 Jurkat proteins with free N-termini and 299 Nt-acetylated proteins in the transamination group (Figure 5.8). A novel finding was the Nt-acetylation of macrophage migration inhibitory factor (MIF_HUMAN/P14174), which had been annotated to only exist in its free

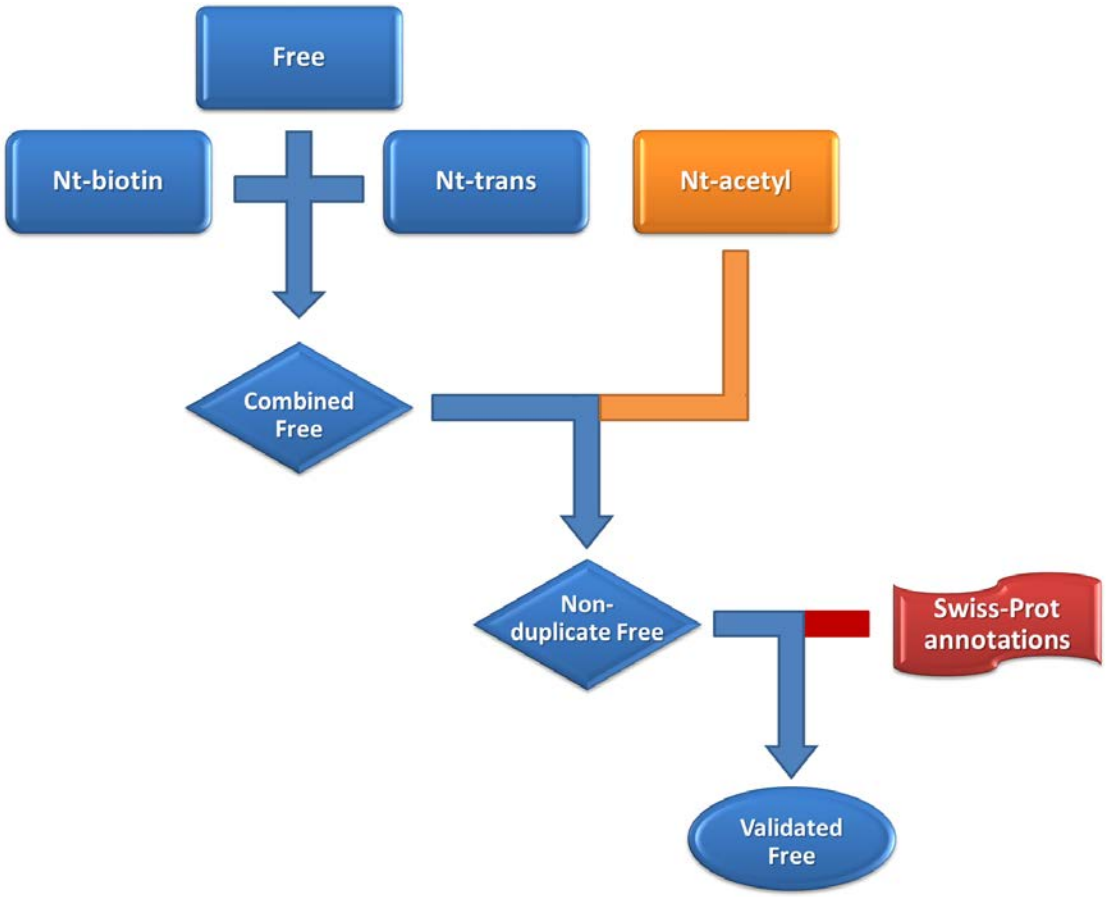


Figure 5.10 Hierarchical strategy used to remove duplicate or conflicting hits in the case of protein N-termini. Nt: N-terminal; trans: transamination; acetyl: acetylation.

form. Overall, there were 379 Jurkat proteins with assigned N-termini (80 free + 299 Nt-acetylated), which was 19 % of the total protein number (2,022) in the transamination group. Free and Nt-acetylated proteins accounted for 21 and 79 % of the proteins with assigned N-termini (or 4 and 15 % of the total proteins), respectively.

In particular, the transamination approach helped to assign the N-termini of 19 Jurkat proteins: 17 in the Nt-trans group and three in the Nt-biotin group (with one overlapping protein hit, see below). A further comparison with the free group showed that only three protein N-termini were uniquely assigned through the transamination approach (i.e. they were absent in the free group). Among them, the N-terminus of tubulin beta-1 chain (TBB1_HUMAN/Q9H4B7) was solely detected in the Nt-trans form, whereas the N-terminus of 60S ribosomal protein L6 (RL6_HUMAN/Q02878) was only detected after transamination and biotin tagging (i.e. exclusive to the Nt-biotin group). The third protein, 60S ribosomal protein L32 (RL32_HUMAN/P62910), was present in both the Nt-trans and Nt-biotin groups as mentioned above. This result suggested that the N-terminus of this protein could be transaminated but the subsequent biotin tagging did not proceed to completion.

Finally, the control and transamination data were merged to compile a complete list of Jurkat protein N-termini identified in the present study. Overall, 2,275 Jurkat proteins were identified and 22 % of them (i.e. 493) were assigned with N-termini (103 free + 390 Nt-acetylated). Among these 493 proteins, 61 % were identified in both the control and transamination groups. In contrast, 23 % of the proteins were unique to the control group and the remaining 16 % were exclusive to the transamination group (Figure 5.11 left). Regarding the state of N-terminal PTMs, 79 % of these proteins were in the Nt-acetyl group and the remaining 21 % possessed free N-termini (Figure 5.11 right). The complete list of the assigned protein N-termini is shown in Appendix 3.

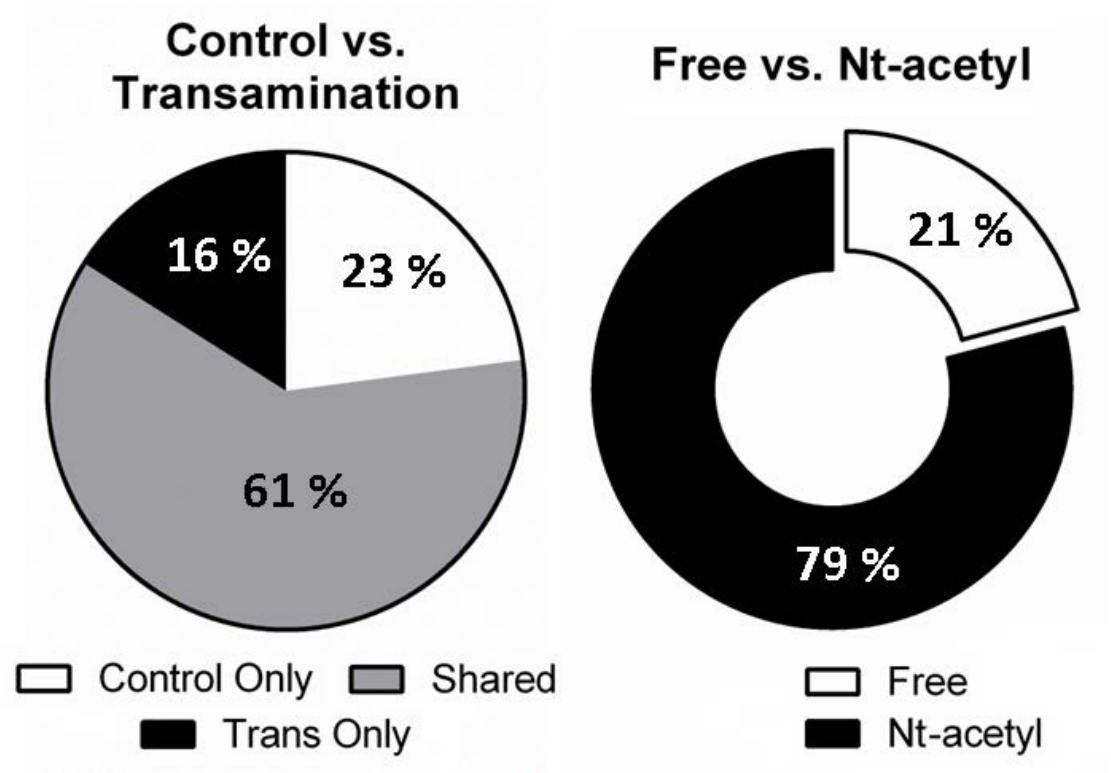


Figure 5.11 (Left) Contribution of the transamination approach to the assignment of overall 493 protein N-termini. (Right) Classification of the 493 assigned protein N-termini according to their N-terminal post-translational modifications (PTMs): "free" or "Nt-acetyl". Trans: transamination; Nt-acetyl: N-terminal acetylation.

The complete dataset of the assigned protein N-termini was then compared with publicly available datasets from other N-terminalomic studies. Subtiligase is a positive selection strategy that has been applied to identify caspase substrates in apoptotic Jurkat T-cells (Mahrus *et al.*, 2008). The resulting data contained not only *neo*-N-termini resulting from caspase cleavage but also free protein N-termini as background. Therefore, it allowed for the comparison of free protein N-termini between the present and Subtiligase studies. With respect to the Subtiligase study, protein N-termini (i.e. peptides starting with the first/second residue in the canonical protein sequences) identified in either control or apoptotic cells were merged to generate a complete dataset of Jurkat proteins with free N-termini.

The complete Subtiligase dataset contained 106 proteins with free N-termini, whereas the present study identified 103 such proteins (see above). Interestingly, these two datasets showed strikingly little overlap: 95 proteins with free N-termini were exclusive to the present study; 98 such proteins were only identified in the Subtiligase study; only eight proteins with free N-termini were shared by both (Figure 5.12). Therefore, the present study and the Subtiligase study identified two distinct sets of free protein N-termini, which might be attributed to undersampling.

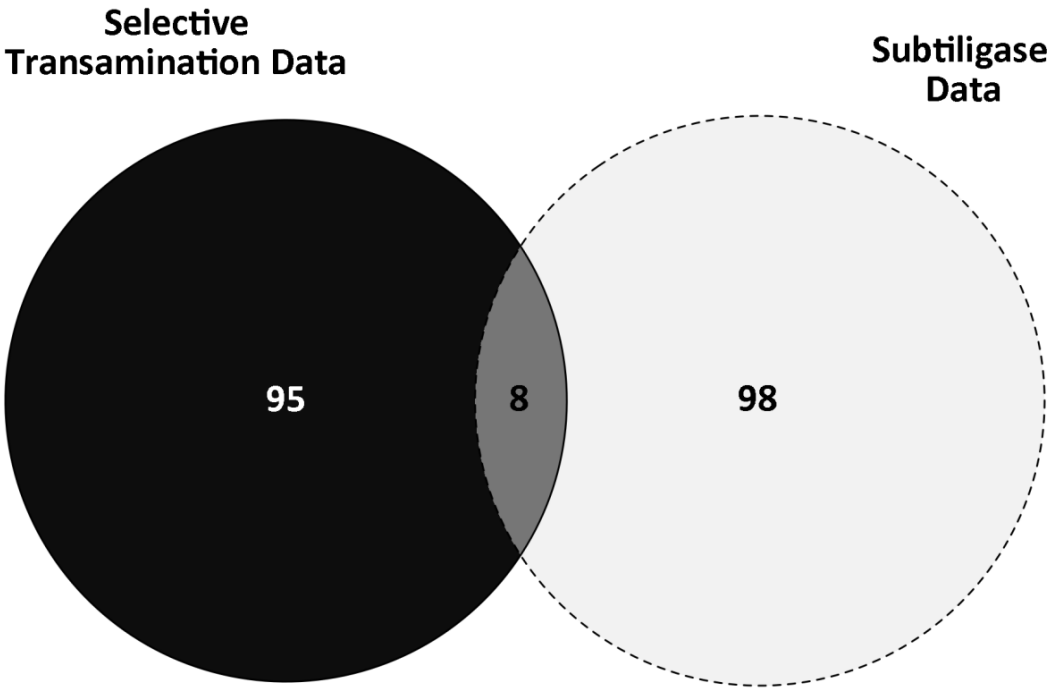


Figure 5.12 Comparison of the identified Jurkat proteins with free N-termini between the selective transamination and Subtiligase datasets. Protein N-termini were extracted from the original Subtiligase data with the following filter: peptide start residue = 1 or 2. The resulting data were then processed to combine the proteins identified in either control or apoptotic Jurkat T-cells.

As mentioned previously, N-terminal COFRADIC is a negative selection strategy that can be employed to identify proteins with either free or blocked N-termini. This strategy has been applied to comprehensively analyse protein N-termini in Jurkat T-cells, and the resulting data are publicly available (Staes et al., 2011). Data from the present study and the N-terminal COFRADIC study were also comparable since both studies could identify free and Nt-acetylated protein N-termini.

As one of the well-developed N-terminalomic strategies, N-terminal COFRADIC assigned 856 protein N-termini in Jurkat T-cells, including 147 proteins with free N-termini and 709 Nt-acetylated proteins. This dataset was compared with the 493 protein N-termini (103 free + 390 Nt-acetylated) assigned in the present study. The comparative analysis showed that 60 % of the protein N-termini (297/493) assigned in this study were absent in the N-terminal COFRADIC dataset (Figure 5.13). A similar trend was also observed when only focusing on free N-termini or the Nt-acetylated ones (Figure 5.14). In conclusion, the present study identified a distinct set of protein N-termini in Jurkat T-cells by comparison with the N-terminal COFRADIC dataset.

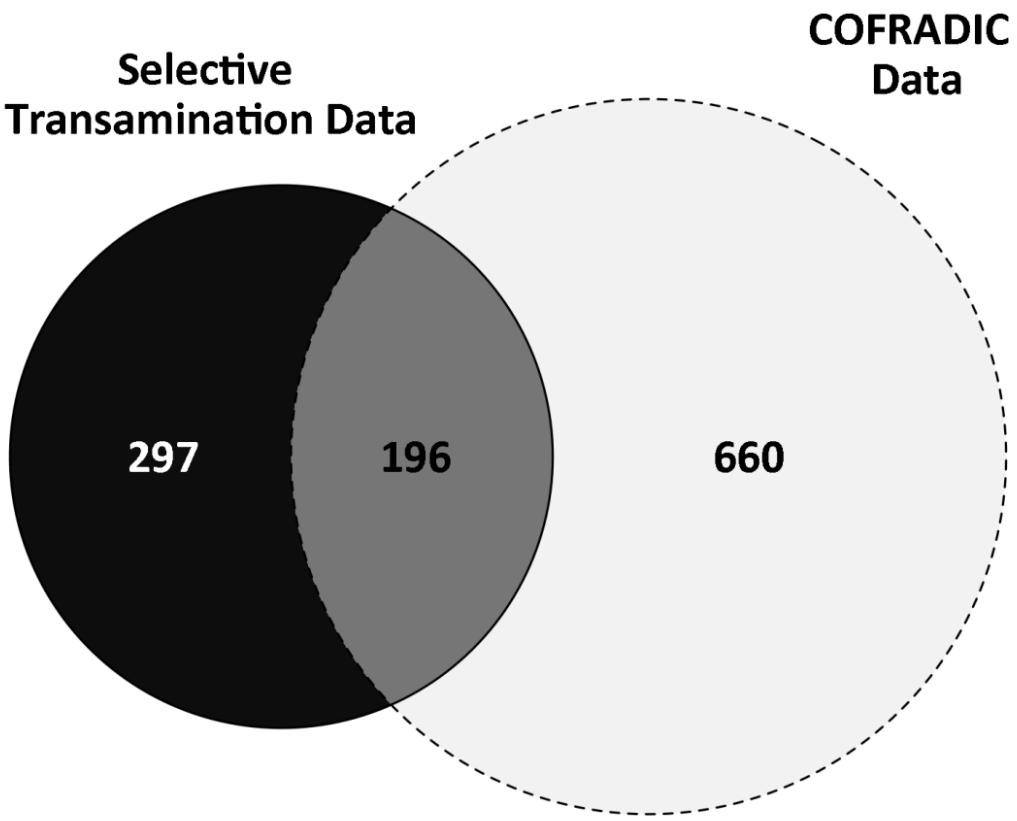


Figure 5.13 Comparison of the identified protein N-termini in Jurkat T-cells between the selective transamination and N-terminal COFRADIC datasets. Protein N-termini were extracted from the original N-terminal COFRADIC data with the following filter: peptide start residue = 1 or 2. COFRADIC: combined fractional diagonal chromatography.

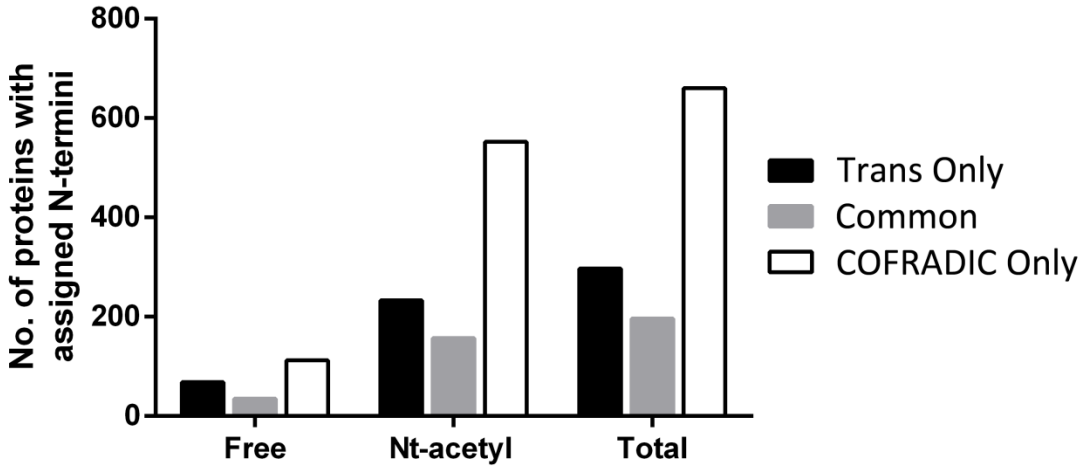


Figure 5.14 Comparison of free, Nt-acetylated, and total protein N-termini assigned in Jurkat T-cells between the selective transamination and N-terminal COFRADIC datasets. "Trans Only": protein N-termini exclusive to the selective transamination dataset; "Common": protein N-termini shared by both the selective transamination and N-terminal COFRADIC datasets; "COFRADIC Only": protein N-termini exclusive to the N-terminal COFRADIC dataset. Protein N-termini were extracted from the original N-terminal COFRADIC data with the following filter: peptide start residue = 1 or 2. COFRADIC: combined fractional diagonal chromatography.

As demonstrated by the above comparisons, none of these N-terminalomic strategies achieved complete coverage of the N-terminal proteome in a single cell line. On the contrary, drastically different sets of protein N-termini were identified using different strategies. This observation seems to support the use of selective transamination, when combined with biotin tagging and affinity purification (AP), to complement other positive selection strategies (e.g. Subtiligase). But even the negative selection strategies (e.g. N-terminal COFRADIC) may benefit from a parallel experiment that employs selective transamination.

5.2.3 Proteomic analysis of Jurkat T-cells using selective transamination

As indicated previously, using the instrumentation available to us, it is routine to identify > 1,000 proteins from unfractionated Jurkat T-cell samples. Without prior fractionation or the use of multiple proteases, the size of this collection is not comparable with those reported by other groups (e.g. Bekker-Jensen *et al.*, 2017). Nevertheless, larger datasets are obtained typically at the cost of longer instrument time. Previously, a large-scale study required three days of instrument time alone to survey the proteome of a cell line (Geiger *et al.*, 2012).

In principle, selective transamination provides a promising opportunity to improve protein identification in Jurkat and other cells without significantly sacrificing the speed of analysis. As described previously, the transamination reaction selectively replaces the N-terminal α-amino group of a protein/peptide with a carbonyl group. Consequently, a modified peptide not only exhibits a -1.03 Da mass shift but also a loss of one positive charge. In addition, transamination alters the behaviour of the modified peptides during reversed-phase high-performance liquid chromatography (RP-HPLC), which likely indicates a change in peptide hydrophobicity. Potentially, such changes in peptide charge states and hydrophobicity could be exploited to achieve orthogonal separation that in turn improves peptide detection and hence protein identification.

A similar concept is adopted by the COFRADIC strategy, which separates the peptides of interest (e.g. N-terminal peptides) from others through shifts in peptide hydrophobicity (Figure 5.15; Gevaert *et al.*, 2003). Although the COFRADIC strategy is elegant and potentially powerful, it is rather labour-intensive, which has hampered its widespread use. In contrast, transamination is a simple reaction but may exert a similar effect on peptide separation.

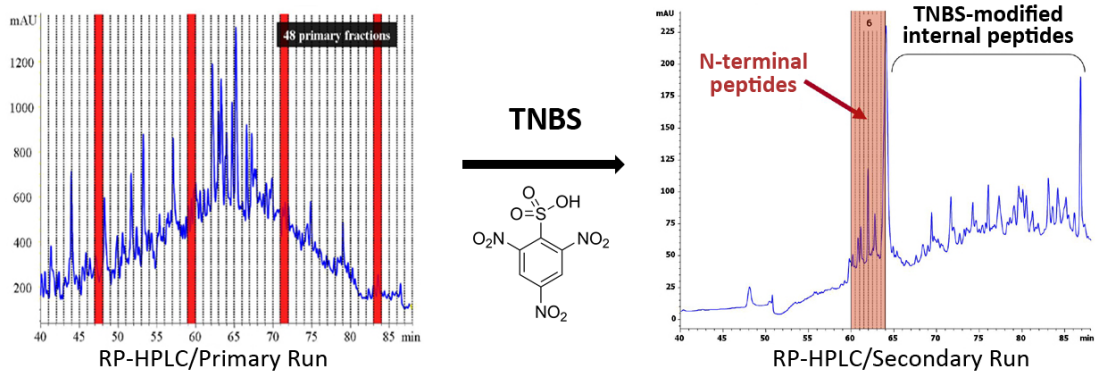


Figure 5.15 Simplified illustration of N-terminal peptide sorting by the COFRADIC strategy (modified from http://penyfan.ugent.be/labo/joelv/Cofradic.html). The reaction with TNBS increases the hydrophobicity of internal peptides, which can be exploited to achieve the negative selection of N-terminal peptides. COFRADIC: combined fractional diagonal chromatography; TNBS: 2,4,6-trinitrobenzenesulphonic acid; RP-HPLC: reversed-phase high-performance liquid chromatography.

Since selective transamination potentially alters peptide charge states and hydrophobicity, it was postulated that transamination at the peptide level would enable the identification of peptides that were otherwise undetectable. Therefore, in the present study transamination was carried out on all tryptic peptides from Jurkat proteins after the protease digestion. The transaminated peptides were directly analysed by LC-MS/MS without further modifications (e.g. biotin tagging). Both the transamination and the corresponding control (i.e. without transamination) groups involved three biological replicates, each with two repeats of spectral acquisition (N = 6). Peptides identified in either group were then pooled for group comparisons. The hypotheses of this study were: I. different sets of tryptic peptides should be identified in the control and transamination groups; and II. combining the two datasets should improve the sequence coverage and hence increase the robustness of protein identification, and possibly also increase the total number of identified proteins.

In either the control or transamination groups, each of the six replicates detected a set of tryptic peptides with high confidence (*E*-value ≤ 0.05). For either group, the tryptic peptides detected by the six replicates were compared by drawing a 6-set Venn diagram (Figure 5.16). The cumulative addition of tryptic peptides by the six control replicates gave rise to a total of 10,388 significant peptides, whereas the total number of significant peptides was only 5,164 in the transamination group (Figure 5.17).

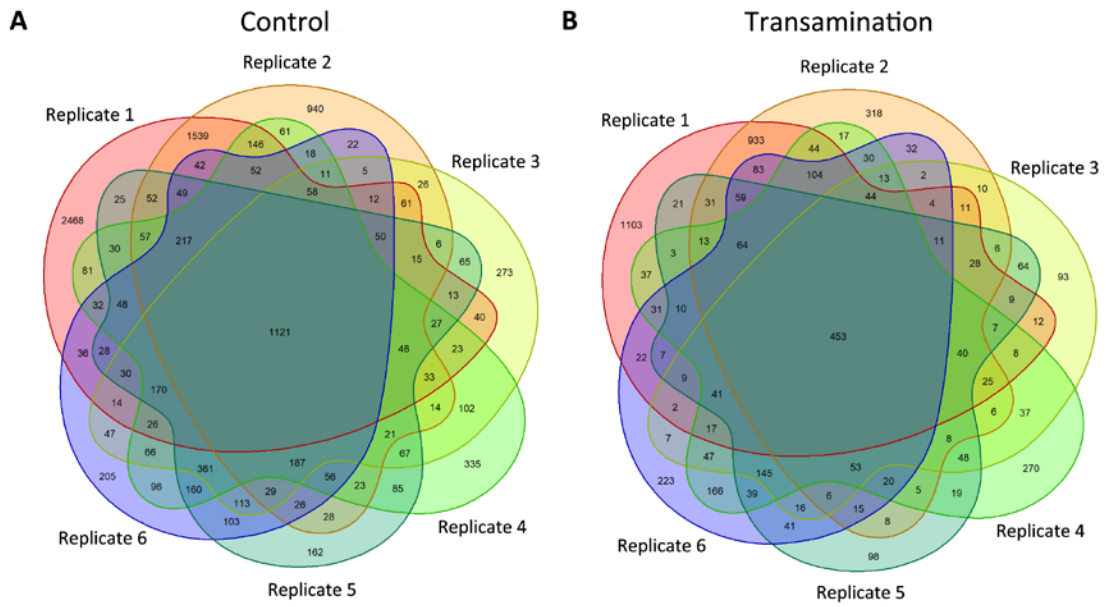


Figure 5.16 Venn diagram comparing the detected peptides between every two out of six replicate analyses in the control (**A**) or transamination (**B**) experiments.

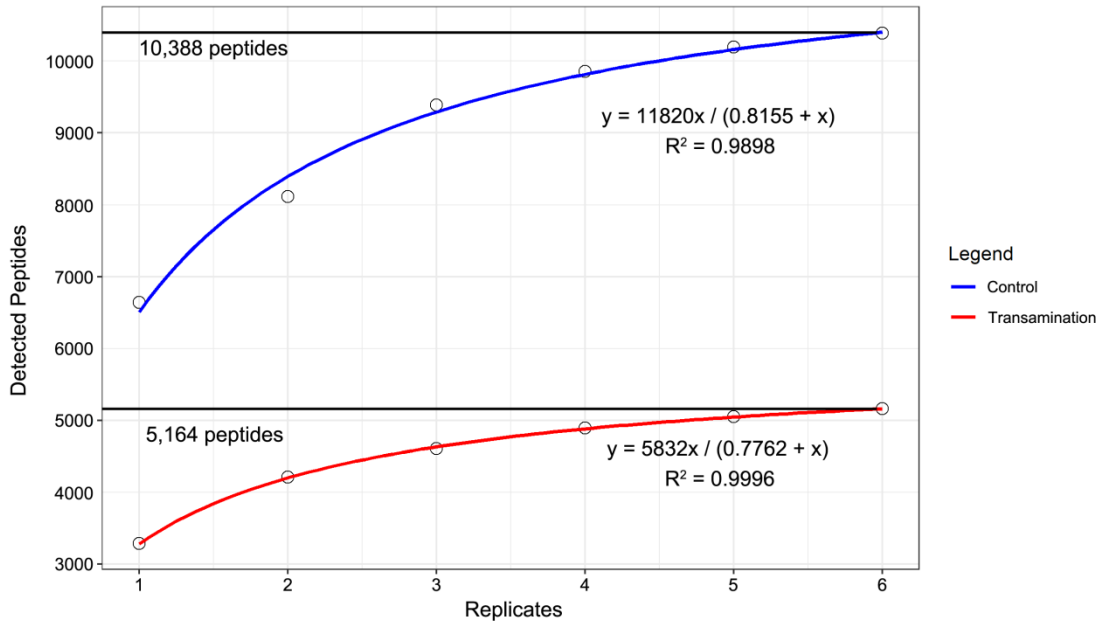


Figure 5.17 Saturation curves of unique tryptic peptides detected by replicate analyses in the control (blue) or transamination (red) experiments. For each group, the curve was fitted through a nonlinear regression analysis in R; in each case, the total peptide number is indicated by a straight line in black.

After combining the data from both the control and transamination groups, a total of 12,834 non-duplicate peptides were reliably detected (Figure 5.18). Among them, 2,718 peptides were shared by both groups. In contrast, there were 7,670 peptides only detected in the control group and 2,446 peptides exclusive to the transamination group. Therefore, selective transamination contributed to a 25 % increase in the number of detected peptides relative to the control data. Moreover, a similar degree of improvement in peptide detection was observed in each pair of the replicate analyses (data not shown). The complete data can be retrieved from the ProteomeXchange Consortium depository (dataset ID: PXD009427).

The effect of selective transamination on peptide separation and detection was further investigated by plotting the distribution of detected peptides over the retention time (*RT*) in RP-HPLC. Compared to the control group (10,388 tryptic peptides), peptides detected in the transamination group (5,164) were more evenly distributed in spite of a lower total number. As a result, a higher proportion of such peptides were detected after ~ 200 minutes (min) in RP-HPLC (Figure 5.19). Furthermore, the shapes of peptide distributions resembled their respective RP-HPLC elution profiles, as revealed by the total ion current chromatograms (TIC) before and after transamination (Figure 5.20). This observation suggested that selective transamination indeed altered peptide behaviour during RP-HPLC, which in turn led to the detection of a different set of peptides by MS/MS. The fact that the transaminated peptides

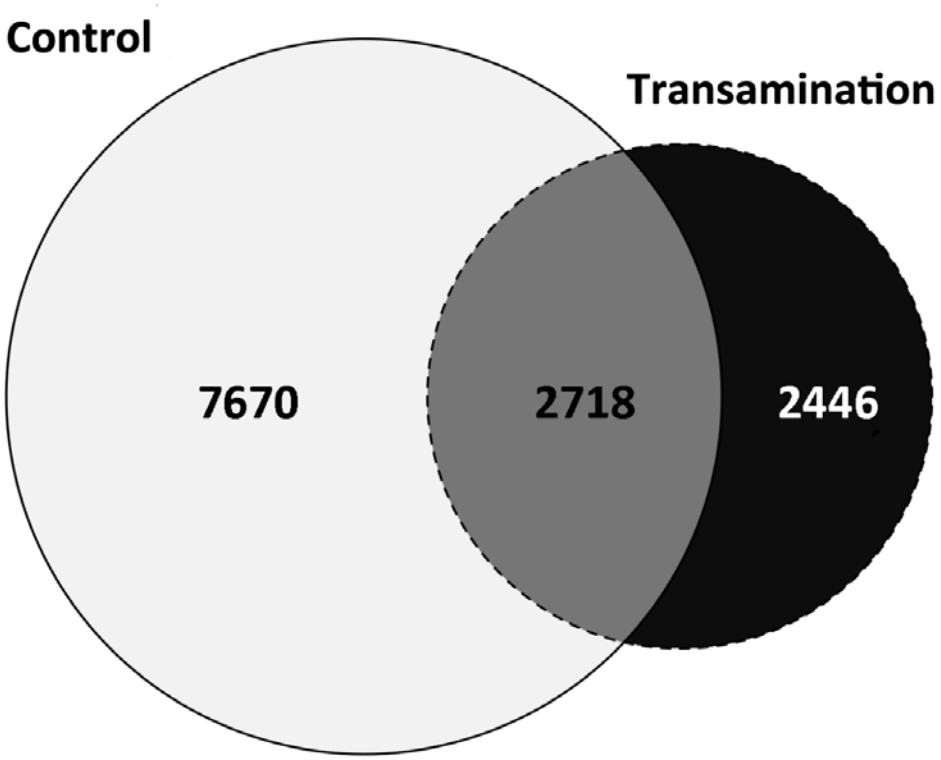


Figure 5.18 Venn diagram summarising the increase in the number of peptides detected by LC-MS/MS due to selective transamination. The control experiment (N = 6) detected 10,388 tryptic peptides from Jurkat proteins, and the transamination experiment contributed to the detection of 2,446 additional peptides. There are 2,718 peptides shared by both groups. LC-MS/MS: liquid chromatography–tandem mass spectrometry; N: number of replicate analyses.

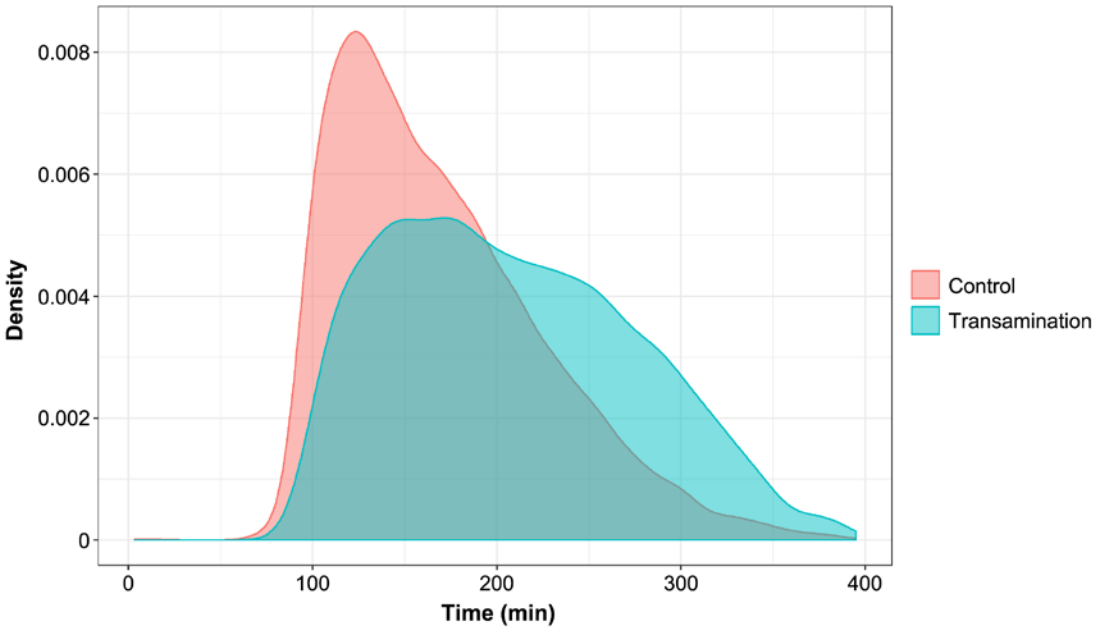


Figure 5.19 Distribution of tryptic peptides detected in the control or transamination experiments. Data are presented as a plot of kernel density estimation. The x-axis represents the retention time (*RT*) in RP-HPLC, whereas the y-axis (density) is an estimate of the peptide counts at a given *RT*. RP-HPLC: reversed-phase high-performance liquid chromatography.

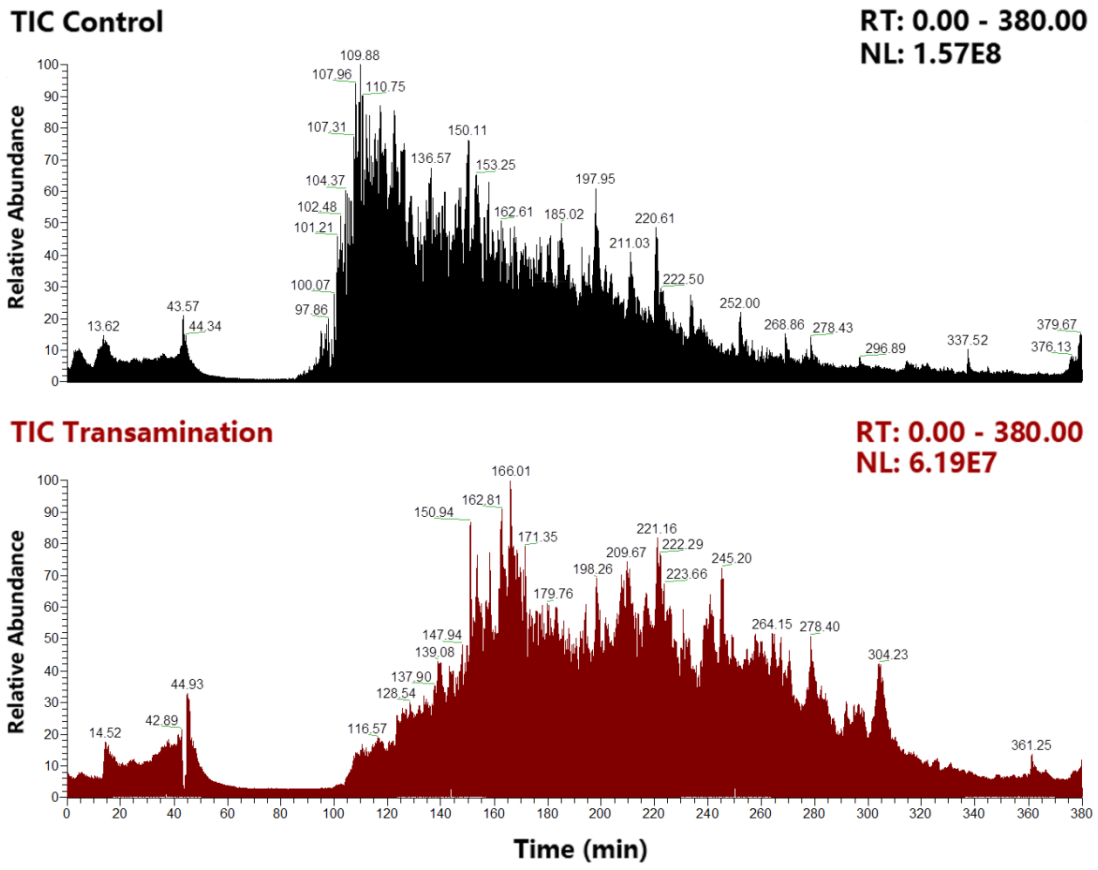


Figure 5.20 Total ion current chromatogram (TIC) of tryptic peptides from Jurkat proteins in the control (black) or transamination (red) groups. *RT*: retention time; NL: normalised intensity level.

eluted more evenly during RP-HPLC might also be advantageous when conventional DDA is used in LC-MS/MS analysis. This is because the average number of co-eluting peptide ions per unit time is likely to be lower following transamination.

Next, the observed effect of selective transamination was also statistically validated using a hierarchical cluster analysis. This statistical analysis computed the degree of dissimilarities (distance) between every two normalised measurements for all six replicates. For both the control and transamination groups, peptides detected by each replicate were indeed clustered within the group and distant from the other group (Figure 5.21). It was therefore concluded that the observed improvement in peptide detection was indeed attributed to the treatment (i.e. transamination), which altered the behaviour of tryptic peptides during RP-HPLC to allow their detection by MS/MS, instead of simply a higher number of replicate analyses.

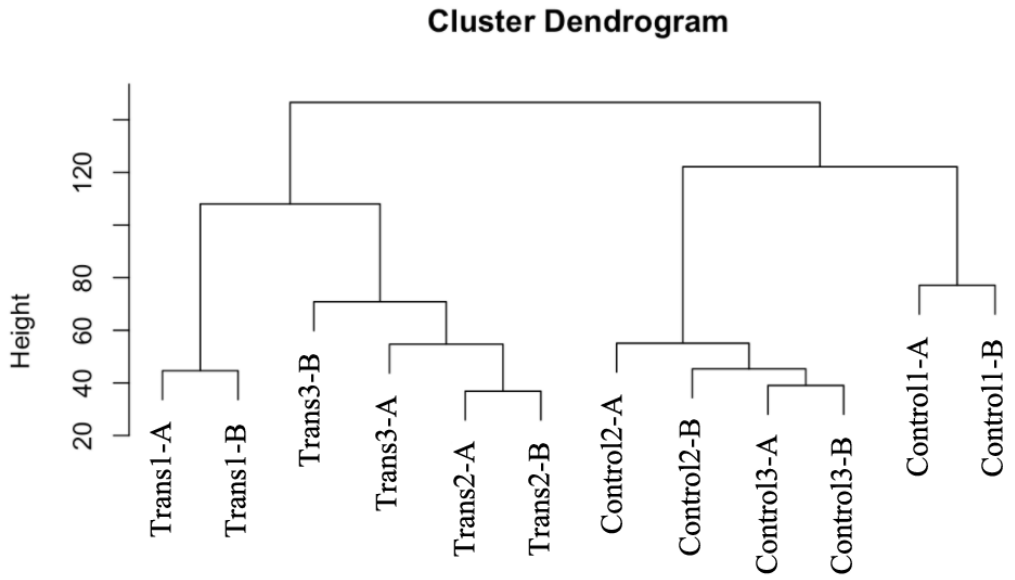


Figure 5.21 Hierarchical cluster analysis of the tryptic peptides identified in the control and transamination groups. For both groups, all six replicates are shown to cluster within the group. A/B denotes two separate technical replicates in each independent biological replication.

The behaviour of peptides during RP-HPLC and electrospray ionisation (ESI) is tightly associated with two physico-chemical properties, isoelectric point (pI) and hydrophobicity (Mirzaei and Regnier, 2006). For instance, the pI value reflects the basicity of a peptide, which has been suggested as a key predictor of peptide charge states in ESI-MS (Liu *et al.*, 2011). Meanwhile, peptide hydrophobicity is correlated with not only the *RT* in RP-HPLC but also the ionisation yield in ESI-MS since this property seems to affect peptide desorption (Mirzaei and Regnier, 2006, Osaka and Takayama, 2014). In view of these arguments, the present study set out to investigate the pI and hydrophobicity of the detected peptides. The hypothesis was that the detection of different peptides between the control and transamination groups could be explained by the changes (if any) in these two properties.

To test this hypothesis, three datasets were extracted from the peptide detection data: control-only (peptides exclusive to the control group), trans-only (peptides solely detected in the transamination group), and common (peptides shared by both groups). For all three datasets, both the theoretical pI and hydrophobicity index of the detected peptides were computed solely on the basis of their amino acid composition. Data analysis was performed using the *Peptides* R package (Osorio *et al.*, 2015).

The results showed a marked difference in peptide pI among the three datasets (control-only, common, and trans-only). As shown in Figure 5.22, the relationship of median pI ($\widetilde{\mathrm{pI}}$) for the three groups is as follows: $\widetilde{\mathrm{pI}}_{\text{control-only}}$ (4.26) < $\widetilde{\mathrm{pI}}_{\text{common}}$ (5.49) < $\widetilde{\mathrm{pI}}_{\text{trans-only}}$ (6.53).

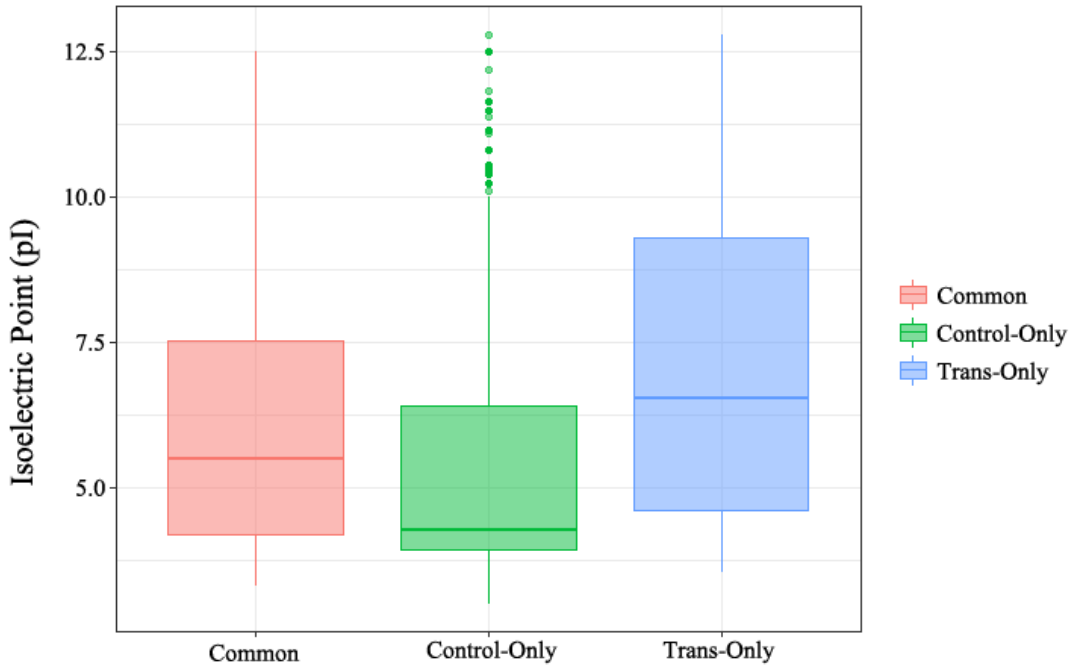


Figure 5.22 Theoretical isoelectric point (pI) of the peptides detected in both the control and transamination groups (common) or in either group (control-only or trans-only). Theoretical pI of each peptide was computed solely on the basis of its amino acid composition. Data are presented as a box plot.

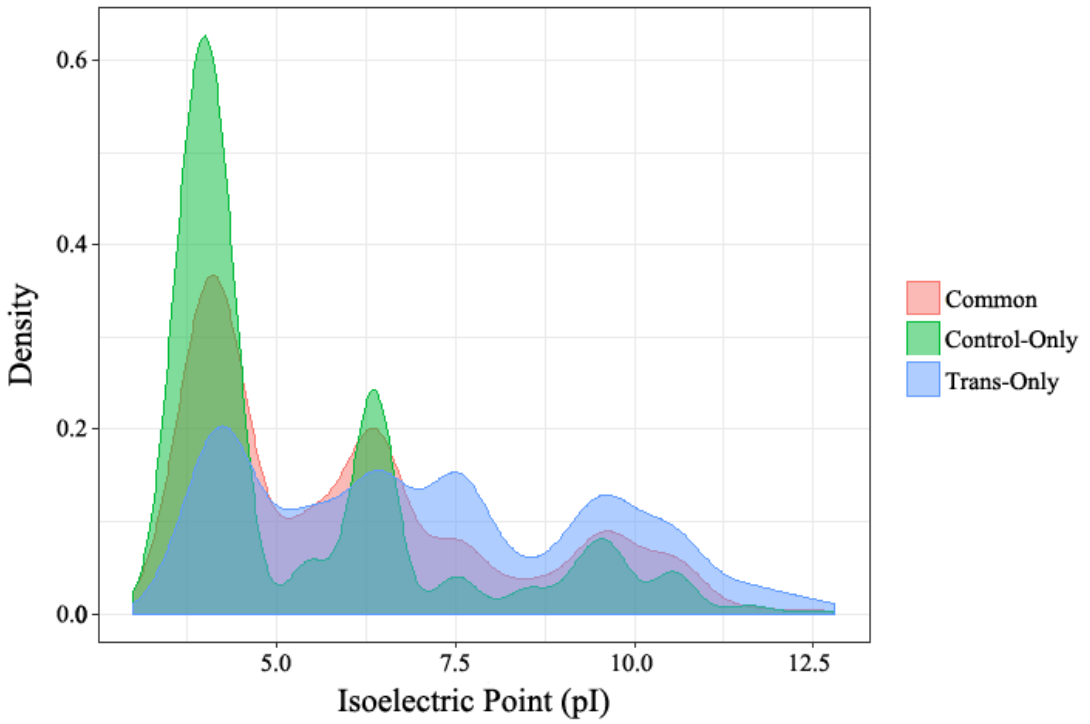


Figure 5.23 Distribution of the theoretical isoelectric point (pI) of the peptides in the common, control-only, and trans-only groups. Data are presented as a plot of kernel density estimation. The x-axis represents the pI range, whereas the y-axis (density) is an estimate of the peptide counts at a given pI value.

Between the control-only and trans-only groups, the aggregate difference in median pI was 2.27. This corresponded to a > 186-fold difference in peptide basicity. In addition, the distribution of peptide pI was also plotted for all three groups (Figure 5.23). Compared to the control-only and common groups, the distribution in the trans-only group clearly shifted towards higher pI. In conclusion, selective transamination allowed the detection of more basic peptides (i.e. with higher pI).

The same datasets were also employed to compute the hydrophobicity index of peptides. Figure 5.24 shows the median hydrophobicity index ($\widetilde{\mathrm{Hphob}}$) for the control-only, trans-only, and common groups: $\widetilde{\mathrm{Hphob}}_{\textbf{control-only}}$ (-0.1143) < $\widetilde{\mathrm{Hphob}}_{\textbf{trans-only}}$ (-0.0031) < $\widetilde{\mathrm{Hphob}}_{\textbf{common}}$ (0.0862). Therefore, peptides solely detected in the control or transamination groups were indistinguishable with respect to the hydrophobicity index. Furthermore, the distribution plot of hydrophobicity index also revealed nearly identical patterns across all three groups (Figure 5.25). These results seemed to support the notion that peptide hydrophobicity was not associated with the detection of different peptides between the control and transamination groups.

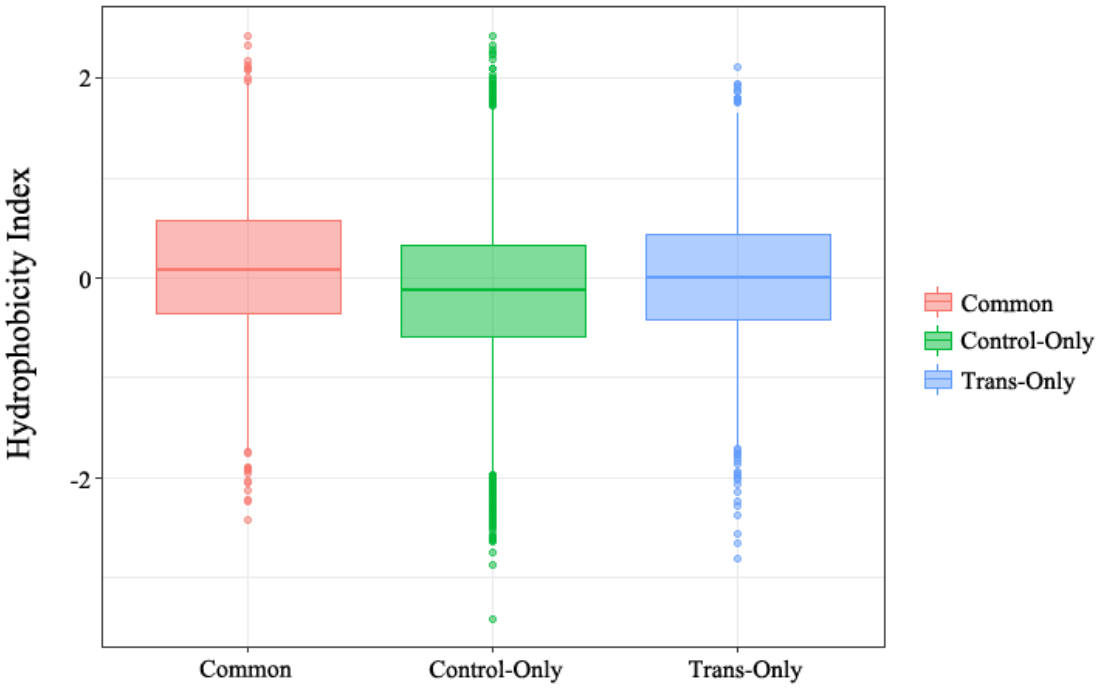


Figure 5.24 Theoretical hydrophobicity index of the peptides detected in both the control and transamination groups (common) or in either group (control-only or trans-only). The hydrophobicity index was calculated using the "Kyte & Doolittle" scale (Kyte and Doolittle, 1982). Data are presented as a box plot.

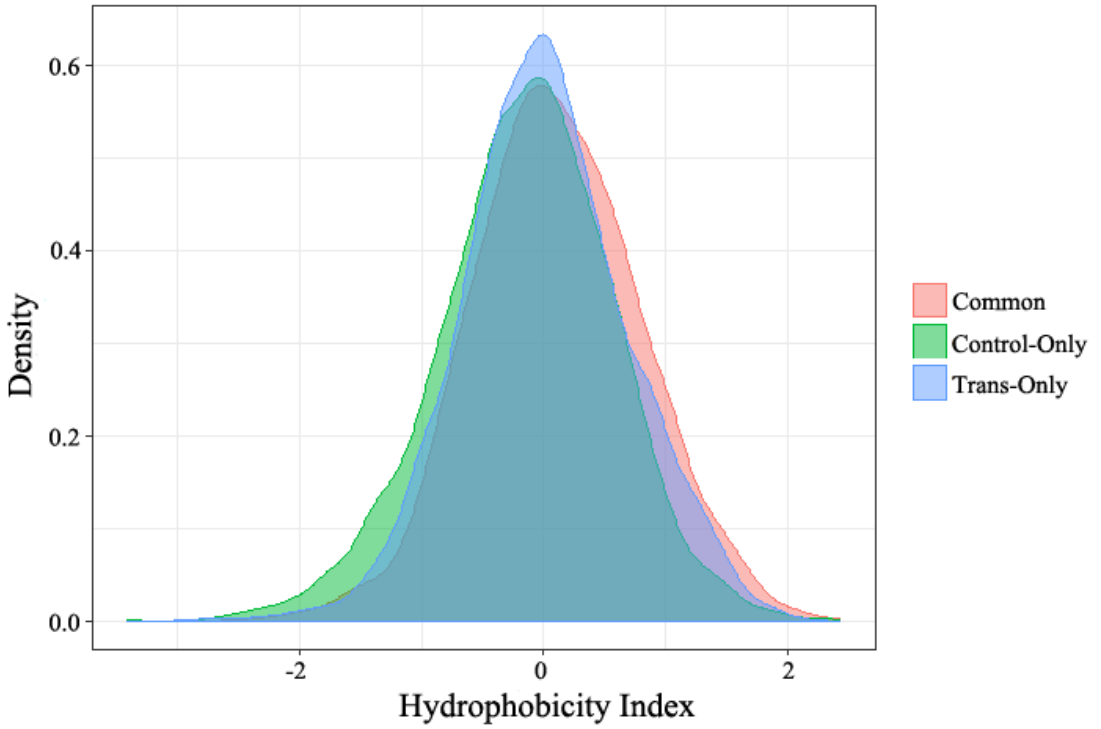


Figure 5.25 Distribution of the theoretical hydrophobicity index of the peptides in the common, control-only, and trans-only groups. Data are presented as a plot of kernel density estimation. The x-axis represents the range of hydrophobicity index, whereas the y-axis (density) is an estimate of the peptide counts at a given hydrophobicity index value.

In addition to the analyses at the peptide level, it was also of interest to investigate the effect of selective transamination on protein identification. Initially, overall 2,963 Jurkat proteins were identified by combining the control (2,533 proteins) and transamination (1,955 proteins) datasets. Since 430 Jurkat proteins were solely identified in the transamination group, this treatment contributed to a 17 % increase in the number of identified proteins relative to the control group (see Appendix 1). However, the total number of identified proteins decreased to 1,594 when the two-peptide rule was applied: 827 proteins were shared by both the control and transamination groups, 664 proteins were exclusive to the control group, and 103 proteins were solely identified in the transamination group (Figure 5.26). As a result, transamination improved the protein identification by 7 % when compared to the 1,491 proteins in the control group. This number is considerably lower than the 25 % improvement in peptide detection, and a plausible explanation is that some of the newly detected peptides were likely mapped to proteins that had already been identified in the control group.

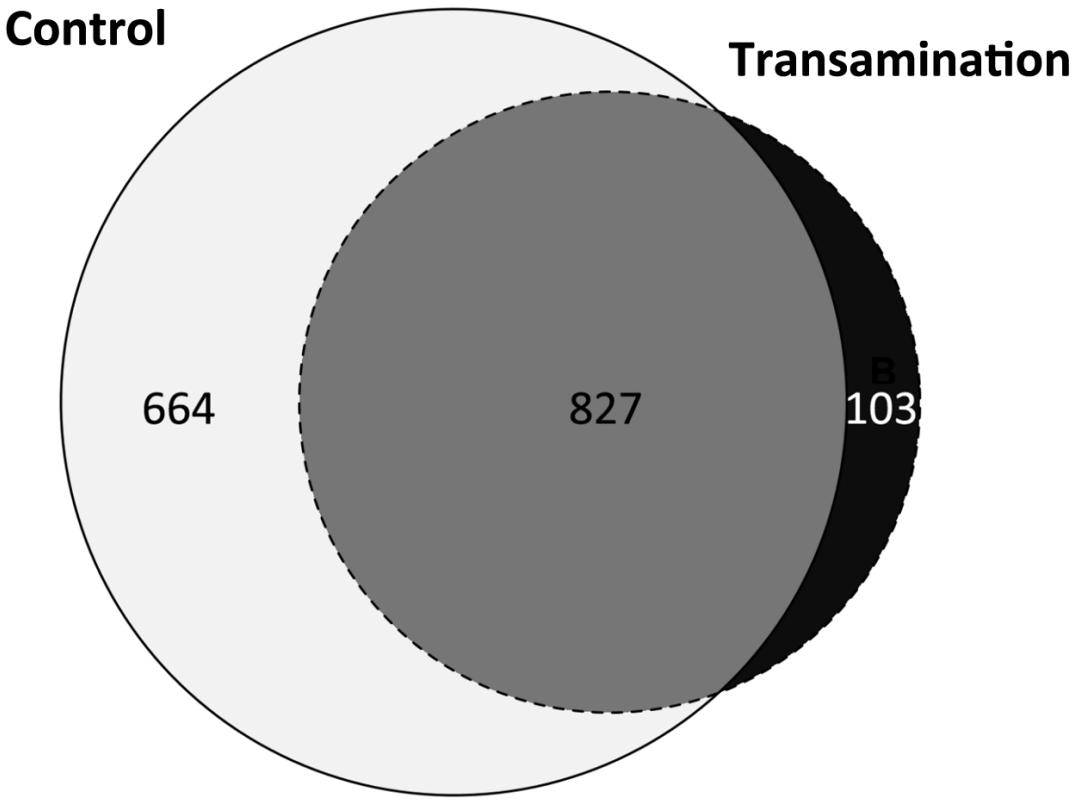


Figure 5.26 Venn diagram summarising the increase in the number of Jurkat proteins identified by LC-MS/MS due to selective transamination. Overall, 1,491 and 930 Jurkat proteins were identified in the control and transamination groups, respectively. In both cases, a protein was identified on the basis of ≥ 2 significant peptide hits (*E*-value ≤ 0.05). The false discovery rate (FDR) was 2.23 %. LC-MS/MS: liquid chromatography–tandem mass spectrometry; *E*-value: peptide expectation value.

The identified Jurkat proteins were analysed for Gene Ontology (GO) term enrichment using the DAVID Bioinformatics Resources v6.8 (Huang *et al.*, 2008). First, the 1,594 proteins identified in the present study were enriched in disease terms including AIDS, infection, and blood cancers when compared to all human proteins in the Swiss-Prot database (Figure 5.27). With respect to protein-protein interactions (PPIs), a diverse range of proteins involved in T-cell activation and general immune response were enriched in this dataset. These proteins include interleukin enhancer-binding factor 2 (Swiss-Prot ID: ILF2_HUMAN/Q12905) and vascular cell adhesion protein 1 (VCAM1_HUMAN/P19320). In addition, two subunits of the T-cell surface glycoprotein CD3 (CD3D_HUMAN/P04234 and CD3E_HUMAN/P07766) were also identified in the present study. These results provided definitive evidence that the identified proteins were indeed extracted from cancer cells and exhibited a T-cell phenotype. The complete list of PPI enrichment is shown in Appendix 4.

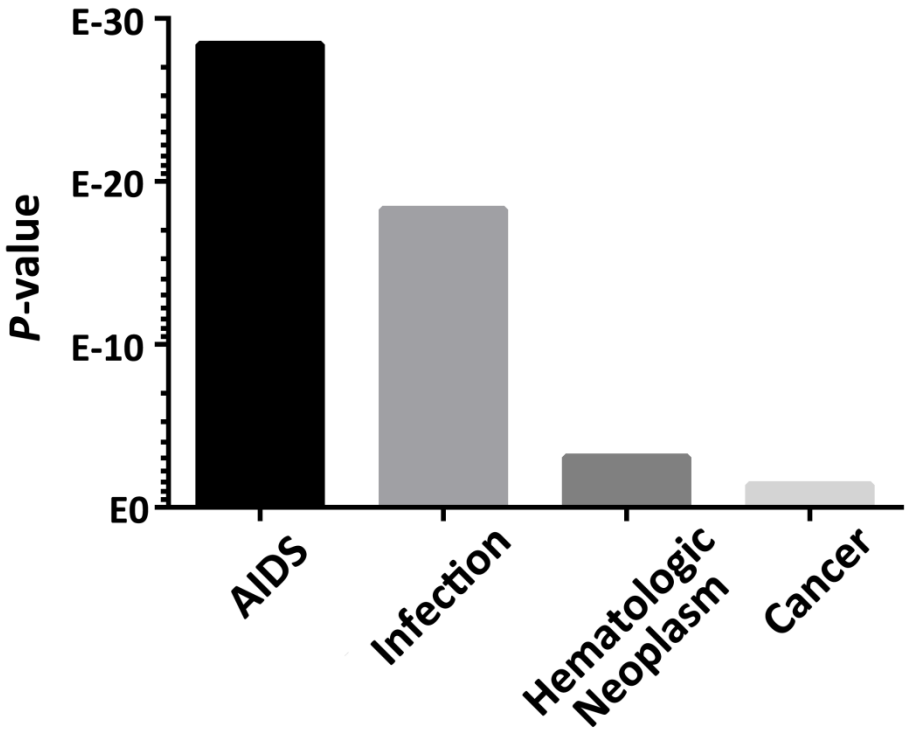


Figure 5.27 Disease term enrichment of all the identified Jurkat proteins relative to total human proteins in the Swiss-Prot database. The x-axis lists the enriched disease terms, whereas the y-axis (*P*-value) represents the expected probability by chance. The enrichment analysis was performed using the DAVID Bioinformatics Resources v6.8 (Huang *et al.*, 2008).

Proteins only identified in either the control or transamination group were then compared with the total identified proteins (1,594) for the enrichment analysis. With respect to GO terms, the 664 proteins in the control-only group were enriched primarily in basic cellular constituents, e.g. integral components of cell membrane and the Golgi apparatus (Figure 5.28A). As expected, protein tyrosine phosphatases were also enriched in resting Jurkat T-cells since they negatively regulate the initiation and propagation of TCR signalling to maintain the resting state (Mustelin and Tasken, 2003). On the other hand, the GO terms enriched in the trans-only group (103) were transport, NADH dehydrogenase (ubiquinone) activity, and kinetochore (Figure 5.28B). Therefore, selective transamination helped to identify a different set of nuclear and mitochondrial proteins in comparison with the control group. However, it was difficult to draw any definitive conclusion from this result due to the lower protein number in the trans-only group.

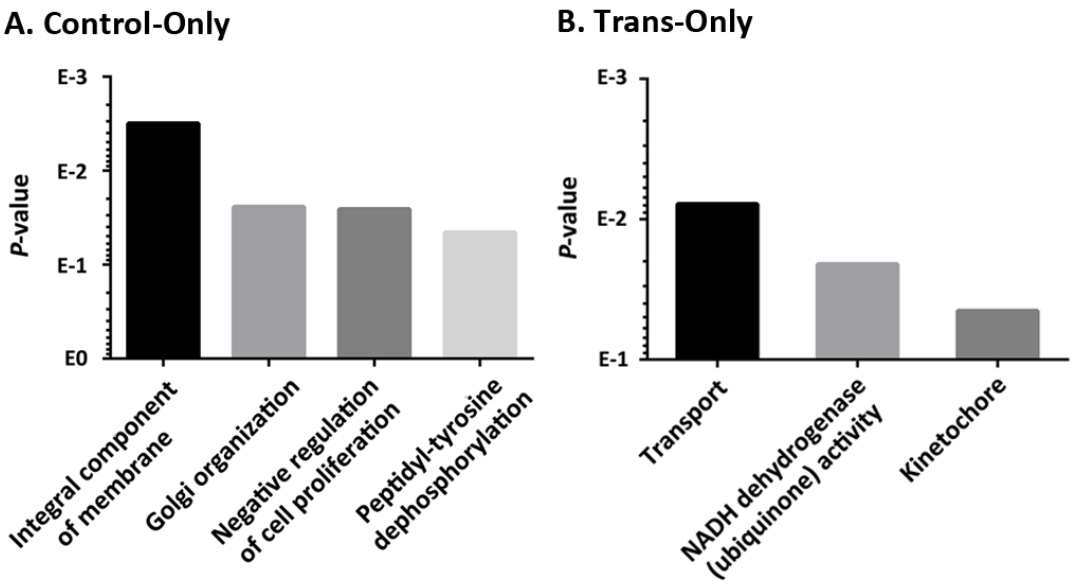


Figure 5.28 Gene Ontology (GO) term enrichment of the proteins only identified in the control (**A**) or transamination (**B**) group. The overall identified proteins (1,594) were set as background. The x-axis lists the enriched GO terms, whereas the y-axis (*P*-value) represents the expected probability by chance. The enrichment analyses were performed using the DAVID Bioinformatics Resources v6.8 (Huang *et al.*, 2008).

In terms of pathway enrichment, the control-only proteins were enriched in mitogen-activated protein kinase (MAPK) signalling, oxidative phosphorylation, and the underlying pathways of the Huntington's disease (HD; Figure 5.29). Interestingly, proteins exclusive to the transamination group were only enriched in the pathways that underlie the Alzheimer's disease (AD). HD and AD are two types of neurodegenerative disorder characterised by the aggregation of misfolded proteins in the brain, and they have several conserved molecular mechanisms (Ross and Poirier, 2004, Flavin *et al.*, 2017). Recently, a transcriptomic study reported the upregulation of many immune-related pathways in both HD and AD (Moss *et al.*, 2017). However, the enrichment of AD or HD in the present study was more likely due to the identification of non-overlapping components of the mitochondrial respiratory chain in either group. For instance, the control-only group contained the subunits of mitochondrial Complex I, III, IV, and V, whereas four additional subunits of Complex I (NADH dehydrogenases) were identified in the trans-only group. These results again supported the hypothesis that different sets of proteins, in this case different subunits of the same protein complex, should be identified in the control and transaminated samples.

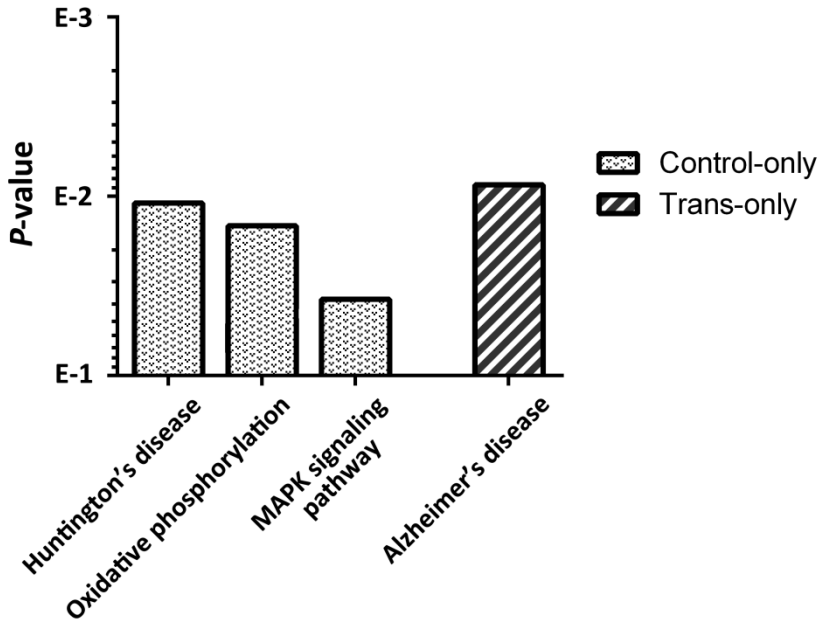


Figure 5.29 Pathway enrichment of the proteins only identified in the control (left) or transamination (right) group. The overall identified proteins (1,594) were set as background. The x-axis lists the enriched pathways, whereas the y-axis (*P*-value) represents the expected probability by chance. The enrichment analyses were performed using the DAVID Bioinformatics Resources v6.8 (Huang *et al.*, 2008).

5.3 Discussion

The N-terminal proteome represents an intriguing target for studying the location, function, and interaction of proteins. However, profiling it proves to be challenging with existing technologies. Even with the state-of-the-art instruments in MS, identification of protein N-termini is often hampered by infrequent detection of N-terminal peptides. It is further complicated by various types of PTMs at protein N-termini in eukaryotic cells. These include NME, Nt-acetylation, and Nt-myristoylation (reviewed in Giglione *et al.*, 2015). Furthermore, proteolytic processing (e.g. signal peptide removal) and alternative translation initiation generate *neo*-N-termini that are previously unknown and hence unannotated in protein databases (Van Damme *et al.*, 2014, Lange and Overall, 2013).

As described previously, several proteomic strategies have been developed to overcome limitations in conventional shotgun approaches and thus improve the identification of protein N-termini/*neo*-N-termini. These include negative selection strategies such as N-TAILS (N-terminal amine isotopic labeling of substrates) and N-terminal COFRADIC, as well as positive selection strategies including N-CLAP and Subtiligase (reviewed in Eckhard *et al.*, 2016). Based on the results described in Chapter 4, the transamination reaction exhibits selectivity for N-terminal α-amino groups over lysine (K) ε-amino groups, and it thus has the potential to be developed as a positive selection strategy for identification of protein N-termini in complex proteomes.

Advancing from the experiments with model peptides and proteins, it was necessary to choose a suitable system with greater complexities for testing the transamination approach regarding its feasibility and efficiency. Jurkat T-lymphocytes were considered as an ideal candidate because they are well characterised and can be conveniently cultured on a large scale in laboratory. Consequently, it is possible to obtain large quantities of complex protein mixtures from this cell line within a relatively short period of time. A three-step study was then carried out to test the transamination approach on this complex proteome: I. biochemical evaluation of transamination and biotin tagging on intact Jurkat proteins; II. global identification of Jurkat protein N-termini (free or Nt-acetylated) through protein-level transamination and biotin tagging; III. shotgun analysis of the Jurkat proteome with the aid of peptide-level transamination.

In the first step, SDS-PAGE and Western blotting analyses demonstrated that Jurkat proteins (DNA-free) were amenable to selective transamination and the subsequent biotin tagging. The Western blotting analysis detected an intense signal of biotin in the treated sample

where Jurkat proteins were subject to both transamination and biotin tagging, but not in the two negative controls without biotin tagging (Figure 5.5). Biotin is naturally bound to five human carboxylases (where it acts as a cofactor in the transfer of CO_2; Wood and Barden, 1977) and perhaps also to histones (Kuroishi *et al.*, 2011). However, the absence of biotin signals in these two negative controls suggests that endogenous protein biotinylation is negligible as compared with the carbonyl-specific biotin tagging after transamination.

Nevertheless, a low level of biotin signal was still detected in a third negative control where Jurkat proteins were directly treated with the biotin tagging reagent without prior transamination. As discussed in Chapter 4, the presence of a biotin signal in this sample likely reflects spontaneous protein carbonylation (a hallmark of oxidative stress; Dalle-Donne *et al.*, 2003). A prompt analysis of protein samples is paramount in avoiding artificial protein carbonylation, as with the cautious use of thiols and metal ion chelators (Luo and Wehr, 2009, Rogowska-Wrzesinska *et al.*, 2014). In addition, hydralazine (a reactive carbonyl scavenger) has been reported to effectively prevent protein carbonylation, but none of the other tested scavengers showed the same effect (Zheng and Bizzozero, 2010). Therefore, addition of this compound during protein extraction may also help to reduce the spontaneous carbonylation. Alternatively, protein carbonyl groups can be blocked with hydrazide or alkoxyamine compounds (e.g. $BnONH_2$) prior to selective transamination.

Another source of biotin signals may be glycoproteins in the extracts of Jurkat T-cells. Conventionally, the primary usage of carbonyl-reactive chemistry is to label glycoproteins after mild periodate oxidation of protein glycans (Weber *et al.*, 1975, Zeng *et al.*, 2009). Therefore, the possibility that glycoproteins were spontaneously oxidised and hence susceptible to carbonyl-specific biotinylation should not be excluded arbitrarily. However, identification of biotinylated glycoproteins is beyond the scope of this study due to the structural complexity of protein glycosylation. In the future, this problem can be largely overcome by either enzymatic or chemical removal of glycans from glycoproteins. For instance, treatment with peptide-*N*-glycosidase F (PNGase F) provides effective removal of *N*-linked glycans (Maley *et al.*, 1989), whereas the reaction with trifluoromethanesulfonic acid (TFMS) has been employed to remove all types of glycans without destroying the protein component (reviewed in Edge, 2003).

Notably, there are several gaps between the results of SDS-PAGE and Western blotting, which lead to the conclusion that some Jurkat protein N-termini are not amenable to selective transamination. The limited scope of transamination can be envisioned on the basis of protein chemistry. For instance, proteins with an N-terminal proline (P) residue have no

reactive α-amino group and thus do not serve as transamination substrates. The limited reaction scope has also been experimentally determined, and will be discussed later in this section. Nevertheless, many Jurkat proteins can still be transaminated and biotin tagged according to the Western blotting results. Therefore, the transamination approach could be employed in a positional proteomic study to globally identify Jurkat protein N-termini.

Positional proteomics is inherently associated with the issue of "one-hit wonders" (Veenstra *et al.*, 2004). In the case of N-terminalomic analysis, protein identification necessarily relies on the assignment of a single peptide (i.e. the N-terminal peptide), which is by definition a one-hit wonder (McDonald *et al.*, 2005). However, Davidson *et al.* (2011) argued that N-terminalomic strategies limit the search space for peptide assignment by providing the positional information, thus greatly increasing the confidence of protein identification. Other researchers have recommended the use of intelligent post-processing tools (Impens *et al.*, 2010) or multiple proteases in parallel experiments (Mommen *et al.*, 2012) to screen the identified one-hit wonders.

However, the issue with one-hit wonders was circumvented in the present study: Jurkat proteins could be identified with ≥ 2 different peptides due to the exclusion of an AP step (i.e. positive selection). Therefore, any protein N-termini assigned were accompanied by other internal peptides and thus did not conform to the definition of one-hit wonders. Such assignments of protein N-termini were also highly reliable given the mass accuracy and resolution of the MS instrument used. Nonetheless, confident assignments were obtained at the cost of undersampling, as some protein N-termini might not be detected by MS due to the overwhelming internal peptides.

Based on the results described in section 5.2.2, this study detected the N-terminal peptides in 493 out of 2,275 Jurkat proteins. It highlights the challenges in profiling the N-terminal proteome as mentioned above: only one fifth of these proteins contained assigned N-termini. Further scrutiny of these protein N-termini has uncovered novel findings: the N-terminal M residue was excised from two proteins, serine--tRNA ligase (Swiss-Prot ID: SYSC_HUMAN) and peptidyl-prolyl cis-trans isomerase (FKBP5_HUMAN); a rare case of Nt-acetylation on a terminal P residue was detected in macrophage migration inhibitory factor (MIF_HUMAN). Such N-terminal PTMs (i.e. NME and Nt-acetylation) have not been reported for these proteins based on existing annotations in the Swiss-Prot database and may be attributed to cell-type specificity in this case. These results highlight the fact that the annotation of N-terminal PTMs is still an ongoing project and will certainly benefit from further generation of proteomic data.

Among the 493 proteins with assigned N-termini, the free-to-acetylation ratio was about 1:4. This ratio is in agreement with the current consensus that the vast majority (> 80 %) of cytosolic proteins in humans are susceptible to Nt-acetylation (Giglione *et al.*, 2015). However, this ratio should be treated with caution as it may vary considerably when the N-termini of the rest 1,782 proteins are also identified. In addition, we have determined that the N-termini of several proteins existed in both free and Nt-acetylated forms. These overlapping protein hits were then removed from the dataset of free N-termini on the ground of incomplete Nt-acetylation, since this PTM is irreversible in nature (Arnesen, 2011). In the human proteome, 9 % of protein N-termini are reported to exist in both free and Nt-acetylated forms (Arnesen *et al.*, 2009). Therefore, these two forms are not mutually exclusive.

Protein transamination and the subsequent biotin tagging proved to be relatively inefficient for assigning protein N-termini. Only 3 out of 103 free protein N-termini were exclusively identified through this approach. Among them, only one N-terminal peptide was assigned solely due to the N-terminal biotin tag. All three N-terminal peptides are derived from highly abundant proteins (e.g. subunits of ribosome or cytoskeleton complex). The scarcity of the biotin-tagged N-termini is probably due to several factors: I. transamination of protein N-termini did not proceed efficiently; II. a single proteoform was transaminated to yield multiple products with different mass shifts, but only the expected product could be assigned by Mascot; III. the abundance of biotin-tagged protein N-termini was further reduced by incomplete biotinylation; IV. without an AP step, MS detection of the transaminated or biotin-tagged protein N-termini was suppressed by internal peptides with higher “flyability”.

A major limitation of the present study is the narrow definition of protein N-termini. Without a local protein database featured with N-terminal annotations, the N-terminus of a protein is automatically defined by the Mascot search engine as the first/second amino acid residue in the canonical sequence of that protein. By this definition, N-terminal biotin tagging and other modifications will only be assigned to the N-terminal M residue or the next one if the position of such modifications is set as “protein N-term”. Consequently, information on two types of N-terminal peptides will be lost. First, true N-termini of mature proteins will not be recognised by Mascot or only treated as internal peptides, since protein maturation often requires proteolytic processing in the N-terminal region. Second, previously unannotated *neo*-N-termini (e.g. due to alternative initiation of translation) will not be discovered.

To partially compensate for the loss of this information, the position of two variable modifications (Nt-trans and Nt-biotin) was further specified to be at the N-termini of semi-tryptic peptides. In this case, these two modifications can be assigned to the N-terminal residue of any peptide from a protein, instead of only the first/second residue in the canonical sequence. A semi-tryptic peptide itself must have a K or arginine (R) residue at its carboxyl (C)-terminus (not further followed by a P residue), but its N-terminus can be any residue. The hypothesis was that this approach should lead to the identification of previously unannotated free N-termini (or *neo*-N-termini), albeit at the expense of a higher false discovery rate (FDR = 3.13 %).

Indeed, a canonical protein N-terminus and 1,214 *neo*-peptides were further identified through this approach (see Appendix 5). Cross-referencing with the Swiss-Prot database validated the canonical N-terminal peptide and three *neo*-peptides as free N-termini of Jurkat proteins (Table 5.3). In the latter cases, the cognate proteins of these *neo*-N-termini have undergone the removal of mitochondrial transit peptides or signal peptides (Hughes *et al.*, 1993, Vaca Jacome *et al.*, 2015). With respect to the effect of biotin tagging, 45 *neo*-peptides were solely identified due to the Nt-biotin modification. Potentially, the 1,214 *neo*-peptides may be derived from processed, stable proteoforms that are physiologically relevant. However, discrimination of these *neo*-peptides from protein degradation intermediates is beyond the scope of this study and requires further analysis.

Table 5.3 List of validated canonical and *neo*-N-termini of Jurkat proteins[a].

Protein Swiss-Prot ID	N-terminal Peptide (Start – End)	N-terminal PTM
Canonical protein N-terminus		
RL40_HUMAN (P62987)	**MQIFVKTLTGK (1 – 11)**	**Nt-trans**
neo-N-termini		
COX5B_HUMAN (P10606)	**ASGGGVPTDEEQATGLER (32 – 49)**	**Nt-biotin**
GLRX5_HUMAN (Q86SX6)	**AGSGAGGGGSAEQLDALVKK (32 – 51)**	**Nt-biotin**
SDF2L_HUMAN (Q9Y4Z0)	**AKTGAELVTCGSVLK (29 – 43)**	**Nt-biotin**

[a] Two variable modifications (Nt-trans and Nt-biotin) were permitted to be at the N-terminus of any semi-tryptic peptide when searching with Mascot. The N-terminus of a semi-tryptic peptide can be any residue but the C-terminus must be either a lysine (K) or arginine (R) residue (not followed by proline, P). False discovery rate (FDR) was 3.13 %. The listed canonical protein N-terminus and *neo*-N-termini were validated by cross-referencing with the online Swiss-Prot database. PTM: post-translational modification; Nt-trans: N-terminal transamination; Nt-biotin: N-terminal tagging with alkoxyamine-PEG_4-biotin or alkoxyamine-PEG_4-SS-PEG_4-biotin.

Two comparative analyses further revealed surprisingly little overlap among the protein N-termini identified by the present and two prior N-terminalomic studies, all of which used the Jurkat cell line. For instance, the present study identified a total of 201 free protein N-termini when combined with a Subtiligase dataset (Mahrus *et al.*, 2008), but only 4 % of them were shared by both datasets. This number only increased to a limited extent (17 %) when comparing the current result with that from a COFRADIC study (Staes *et al.*, 2011). A similar result was shown by the comparison between the Subtiligase and COFRADIC datasets, as only 6 % of the total free N-termini were present in both. Thus, a complete survey of the N-terminal proteome may require collective efforts that each employs a different approach. It should be noted that the strength of Subtiligase lies primarily in the identification of protease substrates, and free protein N-termini were only identified as background in that study. On the other hand, the COFRADIC approach resulted in a much larger collection of protein N-termini (both free and Nt-acetylated), demonstrating the efficacy of negative selection. Finally, such comparative analyses should benefit from integrating the data from other N-terminalomic studies that used the same cell line, e.g. N-CLAP (Xu and Jaffrey, 2010). However, their data are published in a format that prevents reprocessing.

From the results of model peptide/protein experiments, selective transamination is capable of altering peptide *RT* during RP-HPLC even without further derivatisation. The COFRADIC technique employs a similar approach to retain and detect the peptides of interest (e.g. N-terminal peptides; Gevaert *et al.*, 2002, Staes *et al.*, 2011). Accordingly, it was hypothesised that: I. peptide-level transamination would help to detect additional proteolytic peptides by shotgun proteomics; and II. combining the control and transamination datasets should increase the number of identified peptides and proteins. In principle, this treatment is analogous to the use of multiple proteases for improving proteome coverage.

By combining the control and transamination datasets, a total of 12,834 tryptic peptides were detected from six replicate analyses. Among them, 60 % of the peptides were unique to the control group and 20 % were shared by both groups. Peptide transamination contributed to the final 20 % of the peptides (2,446), which would have otherwise been undetected. These figures support the first hypothesis that a different set of tryptic peptides would be detected due to transamination.

Similar to the LC elution profiles extracted from Orbitrap RAW data, the *RT* distributions of the detected peptides varied markedly between the control and transamination groups. For the control group, the detected peptides were more clustered within the 100 – 200 min interval. In contrast, a higher proportion of the peptides were detected after 200 min in the

transamination group. This change in the chromatographic behaviour of peptides (i.e. the shifts in peptide *RT*) is a plausible reason why transamination could help to detect many previously missing peptides. This argument is further supported by the result of a statistical analysis, which showed that the detection of additional peptides was not because of deeper sampling but instead was due to the treatment (i.e. transamination).

In the next step, the present study attempted to determine the possible correlation between transamination and the change in peptide behaviour. Two different physico-chemical properties, peptide pI and hydrophobicity, were investigated. Although a significant difference in peptide hydrophobicity was not identified, there was a substantial change in the pI of the detected peptides following transamination. The result of a computational analysis showed that the peptides in all three groups (control-only, common, and trans-only) had a median pI below 7. However, the transaminated peptides were 17-fold more basic relative to the peptides in the common group, and 186-fold more basic relative to the control-only peptides. The experimental data shown here seem to suggest that: I. LC-MS/MS (positive ESI, DDA mode) more readily detects peptides with lower pI; II. transamination modifies the behaviour of tryptic peptides primarily by lowering their pI, which in turn enables the detection of more basic peptides.

The effect of transamination on peptide pI is possibly attributed to the removal of the N-terminal α-amino group ($pK_a \approx 9.0$; Grimsley *et al.*, 2009, Oregioni *et al.*, 2017). However, interpreting such findings can be difficult: in positive ESI, charges on an ionised peptide are mainly retained by K, R, and histidine (H) residues as well as the N-terminus (i.e. highly basic groups; Krusemark *et al.*, 2009); removing the N-terminal α-amino group may thus be adverse to peptide ionisation/detection. Nevertheless, transamination led to the observed shifts in peptide *RT* during RP-HPLC, which may benefit peptide detection by mitigating the problem of ion suppression, also termed the "matrix effects" (Taylor, 2005). It describes the suppression of (low-abundance) peptide ionisation by co-eluting peptides that are of higher "flyability" (Qian *et al.*, 2006). Since the transaminated peptides no longer co-elute with the suppressive peptides, they are more easily ionised and detected in DDA-mode LC-MS/MS.

The 2,718 peptides shared by both the control and transamination groups were further analysed to unveil any interesting features. We first determined that the enrichment of these peptides was not correlated with their positions in the cognate protein sequences (data not shown). Subsequently, the extent of transamination was assessed using the peptide-spectrum matches (PSMs) retrieved from the transamination dataset. The 2,718 peptides of interest were detected in their native form in the control group but existed in either native or

transaminated forms (or both) in the transamination group. It was calculated that 25 % of the peptides (660) in the transamination group were only detected in the native form. These peptides may thus represent unfavourable substrates for this reaction. In contrast, 17 % of the peptides (450) existed in both native and transaminated forms. Therefore, such peptides were amenable to transamination but the reaction was incomplete. The remaining 59 % (1,608) were exclusively detected in the transaminated form, probably indicating the highly efficient transamination of these peptides.

A comparative analysis was then performed between the 660 peptides that were seemingly refractory to transamination and the 3,977 transaminated peptides. This analysis revealed overrepresentation of the N-terminal threonine (T), valine (V), and H residues in the 660 peptides, as well as moderate enrichment of the P residue at the second position (Figure 5.30). These residues may thus be refractory to transamination or may undergo further reactions. Conversely, leucine (L) is the most underrepresented N-terminal residue, followed by alanine (A) and isoleucine (I). These three residues may represent the optimal substrates for transamination. Taken together, the result of this comparative analysis implies that different N-terminal residues indeed affect the efficiency of transamination.

The above statement is generally in agreement with the limited scope of transamination as experimentally determined using libraries of synthetic peptides: 17 out of the 20 standard amino acid residues can serve as substrates for transamination, but N-terminal H, P, and tryptophan (W) residues do not undergo transamination; the yield of transamination is low when peptide substrates start with a cysteine (C), K, T, or asparagine (N) residue (Papanikos *et al.*, 2001, Gilmore *et al.*, 2006, Sonomura *et al.*, 2009b). Nevertheless, the comparative analysis suggested that the N-terminal V residue was difficult to transaminate, which was not mentioned in the previous reports. Indeed, a further examination showed that 287 out of the 3,977 transaminated peptides start with a V residue but transamination was incomplete for 200 of them. The discrepancy between these results may simply be a result of the higher sample complexity: the present study relied on the use of human protein extracts, which are intrinsically more heterogeneous than the synthetic peptide libraries with respect to the absolute abundance and the distribution of N-terminal residues. An alternative explanation for the discrepancy may be that transamination reduced the “flyability” of peptides with an N-terminal V residue so that they were not detected in the present study.

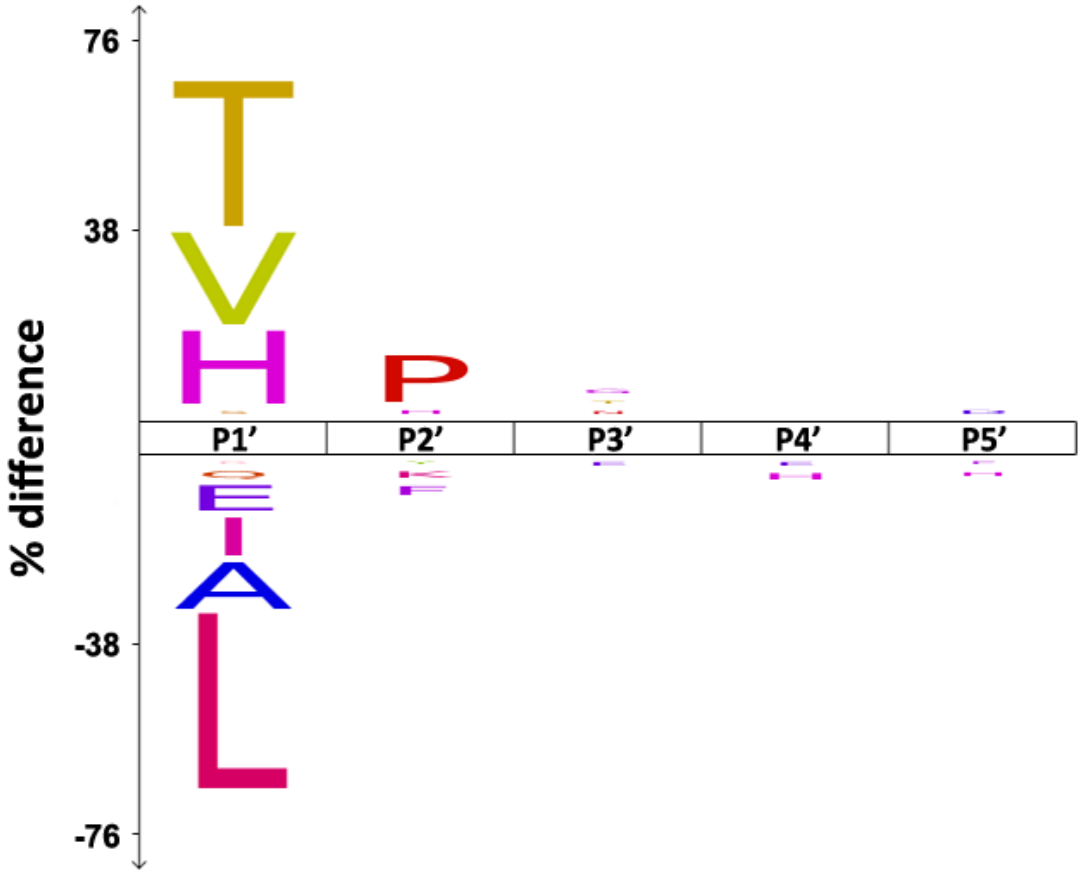


Figure 5.30 Sequence logo depiction of the five amino acid residues (P1' – P5') from the N-terminal end of the 660 peptides that are seemingly refractory to transamination. The y-axis denotes the percentage difference in appearance frequency, relative to the 3,977 peptides that are amenable to transamination (*P*-value = 0.05). The sequence logo was created using iceLogo (Colaert *et al.*, 2009).

At the protein level, there are overall 2,963 proteins identified from the combined dataset, but the removal of one-hit wonders reduced this number to 1,594. Among them, 103 proteins were solely identified from the 2,446 peptides exclusive to the transamination group. Proportionally, there is a substantial difference in the improvements between peptide detection and protein identification: a 25 % increase in peptide detection versus a 7 % increase in protein identification. A plausible explanation is that different peptides detected in the control or transamination groups may contribute to the identification of the same proteins. To test this hypothesis, sequence coverage information of the 827 proteins shared by both groups were retrieved for a comparative analysis. Mean sequence coverage was calculated to be 35.34 and 42.30 % for the control and combined datasets, respectively. Therefore, the 2,446 trans-only peptides contribute to a 7 % increase in the sequence coverage of the 827 shared proteins as well as the identification of 103 additional proteins. It is anticipated that protein identification will be further improved when data-independent acquisition (DIA) is adopted in future studies.

The difference in the improvements between peptide detection and protein identification also owes in large part to the removal of one-hit wonders. As described previously, one-hit wonders refer to proteins that are only identified by a single peptide match, and their validity has been a major concern to the proteomics community (Veenstra *et al.*, 2004). At present, protein identification follows the Paris guidelines for reporting proteomic data (Bradshaw *et al.*, 2006), which enforces the two-peptide rule and rejects one-hit wonders. Consequently, nearly 1,400 proteins identified in the present study were regarded as one-hit wonders and excluded from the final results. This number is roughly equal to that of the proteins identified with high confidence (≥ 2 significant peptides, *E*-value ≤ 0.05). Therefore, highly informative data are potentially lost during this process. This argument is supported by Gupta and Pevzner (2009), who showed that excluding one-hit wonders actually resulted in a lower sensitivity of proteomic studies, and that a large proportion of one-hit wonders were indeed expressed.

Even though the confidence in peptide assignments can be provided by the high resolving power/mass accuracy of state-of-the-art MS instruments and strict FDR control, one-hit wonders still present a major problem at the data analysis step. This "protein inference problem" (Huang *et al.*, 2012) refers to the ambiguity in mapping the identified peptides to their parent proteins due to the existence of one-hit wonders and degenerate peptides (i.e. peptides that are shared by multiple proteins). Several algorithms have been developed to tackle this problem at multiple stages (reviewed in Huang *et al.*, 2012). Recently, an artificial intelligence (AI) method was also introduced to address the challenges in protein inference. This deep-learning technique is termed DeepPep, which employs an artificial neural network to afford robust protein inference by integrating the probability of peptide detection with positional information (Kim *et al.*, 2017). In the future, advances in MS instrumentation and bioinformatics analysis will likely provide an ultimate solution to the problems with one-hit wonders and protein inference.

In conclusion, the present proteomic study employed selective transamination to modify Jurkat protein N-termini for biotin tagging, and to modify the N-termini of tryptic peptides for improved peptide/protein identification. The former experiments identified 490 protein N-termini, with some novel discoveries (e.g. two proteins undergo NME and a third has acetylation at its N-terminus). However, transamination and the subsequent biotin tagging only contributed to the identification of three protein N-termini. The low efficiency reflects incomplete transamination/biotinylation, and highlights the importance of AP. The latter experiments combined the control and peptide transamination data to identify nearly 13,000

peptides or 1,594 Jurkat proteins with high confidence (*E*-value ≤ 0.05). In comparison with the control data, peptide transamination led to a 25 % increase in the number of detected peptides, or a 7 % increase in the number of identified proteins. These improvements were larger than expected and significantly improved proteome sequence coverage. In view of the simplicity of this technique, which we term STONE (Selective Transamination Of N-Ends), it could be implemented widely. Due to its intrinsic limitations, the present study requires further analyses that include: I. amending database and search parameters to discover and validate *neo*-N-termini; II. adopting different data acquisition methods to identify low-abundance proteins; III. employing AI methods to deal with the protein inference problem caused by one-hit wonders.

Chapter 6. Conclusions

The extreme amino (N)-terminus of a protein is generated during protein synthesis, at which point the nascent polypeptide chains start with an N-terminal (formyl-)methionine residue. The N-terminus is highly informative regarding a protein's stability, function, and regulation. The importance of protein N-termini is reflected by the "N-end rule": the degradation signal of a protein has already been planted at its N-terminus from the moment of its birth, linking an N-terminus to its cognate protein's *in vivo* half-life (Varshavsky, 1997). A newly synthesised protein often endures a plethora of co- and post-translational modifications (PTMs) at its N-terminus in order to become fully mature. These PTMs further influence protein stability and activity, as illustrated by another branch of the N-end rule in eukaryotes, "Ac/N-end rule pathway". This pathway describes the regulation of protein half-lives by specific PTMs: N-terminal methionine excision (NME) and N-terminal (Nt)-acetylation (Varshavsky, 2011). Recently, a third branch of the N-end rule was discovered in yeasts, termed the "Pro/N-end rule pathway" (Chen *et al.*, 2017). This pathway refers to the recognition and selective degradation of gluconeogenic enzymes that possess a proline (P) residue at the N-terminus or the second position.

Proteolytic processing represents another major form of PTM. Many newly synthesised proteins require proteolytic processing (e.g. NME) for maturation, thus producing new proteoforms with different N- or carboxyl (C)-termini (i.e. *neo*-termini). In addition, expression of the same gene often gives rise to multiple proteoforms with different N-termini, owing to genetic mutations, alternative splicing, alternative translation initiation, and so on (Lange *et al.*, 2014b). All these events contribute to a greater complexity of the proteome: the apparent N-terminus of a protein does not always conform to that predicted by bioinformatics analysis of the genome. Currently, an emerging subfield of proteomics, termed "N-terminalomics", attempts to resolve this issue. The primary objectives of N-terminalomics are to determine the exact N-termini of all proteoforms in a biological system and the biological relevance of such N-termini. In addition, N-terminalomic strategies are widely adopted in efforts to identify proteolytic processing events and protease substrates.

N-terminalomic strategies are divided into two main categories: positive and negative selection. The former involves direct enrichment of protein N-termini through chemical reactions that differentiate between the α-amino group at protein N-terminus and ε-amino groups on the side chain of lysine (K) residues. In contrast, the latter requires the removal of peptides other than protein N-termini (i.e. internal peptides) and does not require

discrimination between α- and ε-amino groups (Agard and Wells, 2009). This thesis describes the efforts to critically evaluate and improve an existing negative selection strategy, which relies on the use of *N*-hydroxysuccinimide (NHS)-activated Sepharose to remove internal peptides (McDonald *et al.*, 2005). It is referred to as the "NHS-Sepharose" approach.

This work has also focused on a positive selection strategy called "Selective Transamination Of N-Ends (STONE)". This approach is based on a chemical reaction largely developed by Dixon (1964). Two major advantages of STONE are the simplicity to implement this approach and the introduction of a reactive carbonyl group, which greatly facilitates further manipulation. The following conclusions can be drawn from this study:

- The NHS-Sepharose approach efficiently blocked both protein N-termini and the side chain of K residues; however, the efficiency of internal peptide scavenging was not satisfactory due to the presence of specific internal peptides after the negative selection.
- Systematic refinement of this approach resulted in complete removal of internal peptides from the N-terminal peptide of chicken egg-white lysozyme (lysozyme C) or bovine serum albumin (BSA); in principle, these refinements should be applicable to many proteins, thus allowing the implementation of this approach to study complex proteomes.
- The STONE approach efficiently converted the N-terminus of various peptides and proteins to a reactive carbonyl group, which allowed further attachment of a carbonyl-reactive biotin tag to the N-terminal residue of these peptides/proteins.
- It was also possible to purify such biotin-tagged model proteins (lysozyme C and BSA) through avidin-biotin interaction.
- Shotgun proteomics of Jurkat T-cells identified several novel N-terminal proteoforms.
- STONE helped to further identify > 1,000 putative *neo*-N-termini in the Jurkat proteome, with 45 of such *neo*-peptides solely identified by the carbonyl-specific biotin tag.
- STONE complemented standard shotgun proteomics with a 25 % increase in the number of detected peptides, which in turn contributed to a 7 % increase in Jurkat protein identification and 7 % higher sequence coverage of the proteins identified in both the standard shotgun and STONE experiments.

In addition, the present work led to a better understanding of these N-terminalomic approaches. With respect to NHS-Sepharose, we have determined that the reactivity of the amine scavenger (i.e. NHS-activated Sepharose) was a major bottleneck in this approach.

Indeed, we learned that this limitation was a driving force behind the development of some recent N-terminalomic approaches (personal communication). Prolonged treatment with an increased amount of NHS-activated Sepharose achieved complete removal of internal peptides of model proteins, but at the expense of a more complex protocol, greater reagent consumption, and potential sample loss. Meanwhile, blocking of primary amines with *N*-succinimidyl *S*-acetylthioacetate (SATA) requires further optimisation since its efficiency was not comparable to amine acetylation with NHS-acetate. The current protocol of shotgun proteomics is not compatible with citraconic anhydride (CA) modification because the modified peptides will be exposed to a low-pH environment that readily dissociates the amide linkage.

Regarding the STONE approach, the reaction products exhibited a ~ 6-minute increase in the retention time (*RT*) during reversed-phase high-performance liquid chromatography (RP-HPLC). Furthermore, shotgun proteomics of Jurkat T-cells revealed that STONE-modified tryptic peptides eluted more evenly during RP-HPLC. This is possibly attributed to a loss of charges that altered peptide isoelectric points (pI). The reaction scope was also briefly examined through the shotgun analyses: tryptic peptides with an N-terminal threonine (T), valine (V), or histidine (H) residue were less susceptible to transamination. In contrast, peptides starting with a leucine (L), isoleucine (I), or alanine (A) residue were transaminated with very high efficiency. This finding is largely in agreement with that reported by Sonomura *et al.* (2009b), but the impact of the N-terminal V residue requires further analysis. Also in keeping with the previous study, formation of the correct transamination product was accompanied by some minor side reactions, including decarboxylation and glyoxylate addition. It remains to be seen if these can be further reduced or abolished by optimisation of the STONE approach. Finally, it was determined that the carbonyl-specific biotinylation did not efficiently tag all the STONE-modified proteins/peptides, hindering the affinity purification (AP) of biotinylated N-terminal peptides.

The present study also reveals that N-terminal annotations in human protein databases are currently incomplete. For example, the STONE experiments identified three endogenously modified protein N-termini that were previously unregistered. This study also identified > 1,000 putative *neo*-N-termini that might represent stable proteolytic products. These results imply that database annotations of protein N-termini may be incomplete for any organism, due to complications in gene expression and subsequent PTMs. Complete N-terminal annotations thus require unremitting endeavours to generate new proteomic data, with a significant contribution from N-terminalomic strategies. Since these methods have their

respective strengths and weaknesses (Table 6.1), it would be wise to choose an appropriate method that serves a specific purpose (e.g. profiling protein N-termini or identifying protease substrates). Alternatively, parallel experiments using different strategies can be performed to increase the number of N-terminal assignments. Also, such assignments are obviously more reliable if the same peptide is independently identified by two or more orthogonal methods.

As discussed in previous chapters, the STONE approach provides a new option for N-terminalomic studies. As a positive selection approach, STONE possesses unique advantages that well complement other approaches in the same category. Compared to N-CLAP (N-terminalomics by chemical labeling of the α-amine of proteins) and *O*-methylisourea, STONE does not involve modification of ε-amino groups on K side chains. The Subtiligase approach requires a proprietary enzyme and a large amount of samples to start with, whereas STONE employs simple and inexpensive reagents to achieve N-terminal tagging. Conversely, N-CLAP surpasses the STONE approach (in its current state) with a better-defined chemistry. The Subtiligase approach is superior over STONE owing to its single-step biotinylation of protein N-termini. Consequently, STONE may be performed to validate protein N-termini that are assigned by other positive selection approaches. Traditionally, protein N-termini identified through positive selection are questionable due to the problem of "one-hit wonders", i.e. the protein assignments necessarily rely on the identification of a single peptide.

In formal terms, there are two problems associated with protein identifications based on single peptides. The first is that peptide might be erroneously identified, i.e. the tandem mass spectra may be incorrectly matched to a peptide sequence. In principle, such errors should diminish as MS technology improves in terms of mass accuracy and precision. In the short term, this problem can be overcome if the protein N-termini are independently identified by two or more orthogonal methods. The second problem, known as the protein inference problem, is more difficult. In such cases, the peptide is correctly identified but it is impossible to assign it to a single protein because it is degenerate, i.e. the same peptide can be produced by multiple proteins (Huang *et al.*, 2012). Although solving this problem is outside the scope of the present work, it is worth noting that bioinformatics approaches, e.g. using artificial intelligence (AI) strategies, may eventually resolve the issue (Kim *et al.*, 2017). Alternatively, top-down MS strategies may be helpful since they directly acquire sequence and structural information from intact proteins. But much remains to be done before these can be routinely implemented.

Table 6.1 Strengths and weakness of major N-terminalomic approaches[a].

N-terminalomics	Strength	Weakness
COFRADIC ChaFRADIC	Widely available and inexpensive reagents, retains modified protein N-termini	Instrument intensive, require highly skilled personnel, potential sample loss
N-TAILS	Versatile and easy to use, lower demand for MS time, retains modified protein N-termini	Relies on specific amine-scavenging polymers (limited availability), potential sample loss during solid-phase extraction
Subtiligase	Absolute specificity towards the N-terminal α-amine, direct enrichment leads to less false-positives, ideal for identifying *neo*-N-termini due to proteolytic processing	Relies on a proprietary enzyme with limited availability, requires a large amount of protein samples, limited reaction scope leads to false-negatives, unable to identify modified protein N-termini
N-CLAP	Widely available and inexpensive reagents, direct enrichment leads to less false-positives, ideal for identifying *neo*-N-termini due to proteolytic processing	Complicated chemical modifications lead to potential sample loss, difficult to identify modified protein N-termini
STONE	Desired specificity towards the N-terminal α-amine, widely available and inexpensive reagents, well-suited to peptide detection by shotgun proteomics	Potential sample loss during sample handling, produces false-negatives due to limited reaction scope and side reactions, difficult to identify modified protein N-termini

[a] COFRADIC: combined fractional diagonal chromatography; ChaFRADIC: charge-based fractional diagonal chromatography; N-TAILS: N-terminal amine isotopic labeling of substrates; N-CLAP: N-terminalomics by chemical labeling of the α-amine of proteins; STONE: selective transamination of N-ends; MS: mass spectrometry.

A major limitation of all negative selection approaches is the blocking of both α- and ε-amino groups, which potentially destroys protease recognition sites. It may also lead to large-scale protein precipitation and elongated peptides, which adversely impact peptide yields and identification. Since STONE does not modify ε-amino groups, in theory it is ideal for identifying *in vitro* protease substrates and minimising the peptide loss. Furthermore, this approach is less instrument-intensive in comparison with N-terminal COFRADIC and

ChaFRADIC. Finally, STONE is highly versatile in terms of potential applications. In addition to the proposed usage as a positive selection approach, STONE can also be implemented in the negative mode to enrich and identify blocked protein N-termini (Sonomura *et al.*, 2009a). In principle, it can be combined with N-terminal COFRADIC or ChaFRADIC as well. Similar to these approaches, STONE prolongs the *RT* of internal peptides during RP-HPLC and removes a positive charge from each internal peptide, whereas blocked protein N-termini are unaffected. Apart from complementing other N-terminalomic approaches, STONE in its simplest form also increases proteome coverage when combined with shotgun proteomics.

With respect to the NHS-Sepharose approach, future research should focus on improving the efficiency of free amine scavenging that ultimately leads to improved enrichment of protein N-termini. Recently, Yeom *et al.* (2017) described a negative selection protocol built upon the NHS-Sepharose approach (referred to as "N-terminal peptides enrichment on the filter" or "Nrich"). It combined the removal of internal peptides by NHS-activated Sepharose with alternative primary amine blocking (see below) and filtered aided sample preparation (FASP). Iterative interrogations of the acquired MS/MS data using either Swiss-Prot or a customised N-terminal database reported a 50 – 80 % enrichment of protein N-termini. This result implies that compartmentalisation of different chemical reactions reduces the interference from internal peptides.

The Nrich approach employed two parallel reactions, propionation or isotopic acetylation, to block all primary amines. Both reactions enabled discrimination between endogenously modified and free protein N-termini. Each reaction was then followed by either tryptic or Glu-C digestion. As a result, nearly 5,000 protein N-termini were identified with high confidence since they were present in at least two (out of a total of four) independent experiments. In view of this report, we argue that the use of orthogonal methods will greatly increase the confidence of peptide assignments. Such orthogonal methods include not only the implementation of multiple N-terminalomic strategies, but also isotopic/isobaric labelling and parallel protease digestions. Among them, the use of multiple proteases has already been shown to improve the identification of low-abundance proteins in a complex proteome and to increase the protein sequence coverage by ~ 200 % (Swaney *et al.*, 2010).

The simplicity and versatility of STONE should permit its use as an orthogonal method to complement both positive and negative selection approaches to identify protein N-termini or protease substrates. However, a premise of such potential applications is the optimal performance of this approach. In its current state, STONE exhibits varied reaction efficiencies depending on different N-terminal residues. Even efficiently transaminated

proteins or peptides may further yield multiple side products, which could potentially exist in higher abundances than the anticipated transamination product. Consequently, future studies should focus on systematic refinement of the experimental conditions. For instance, addition of ethylenediaminetetraacetic acid (EDTA) has been suggested to stop the reaction and suppress the formation of side products (Sarkar *et al.*, 1978).

Once optimised, STONE can be combined with relative (e.g. stable isotope labeling with amino acids in cell culture, SILAC) and absolute quantitative techniques such as selected reaction monitoring (SRM) or sequential window acquisition of all theoretical fragment-ion spectra (SWATH) to measure the abundance of identified protein N-termini. In the future, it is also desirable to invent one-step (or one-pot) tagging reactions based upon the STONE method. This type of strategy is exemplified by the N-terminal modification with 2-pyridinecarboxaldehyde (2PCA; MacDonald *et al.*, 2015) and metal-free click chemistry (Ning *et al.*, 2010). Such reactions will undoubtedly improve tagging efficiency, reduce sample loss, and eliminate the chance of nonspecific tagging. In conclusion, the simplicity and versatility of STONE and its potential derivatives may drastically advance research in positional proteomics and degradomics.

Bibliography

Abbas, A. K., Lichtman, A. H. & Pillai, S. 2015. *Cellular and molecular immunology*, 8th ed.

Abbas, A. K., Lichtman, A. H. & Pillai, S. 2018. *Cellular and molecular immunology*, 9th ed.

Abraham, R. T. & Weiss, A. 2004. Jurkat T cells and development of the T-cell receptor signalling paradigm. *Nat Rev Immunol,* 4**,** 301-8.

Adams, P., Fowler, R., Howell, G., Kinsella, N., Skipp, P., Coote, P. & O'Connor, C. D. 1999. Defining protease specificity with proteomics: a protease with a dibasic amino acid recognition motif is regulated by a two-component signal transduction system in *Salmonella*. *Electrophoresis,* 20**,** 2241-2247.

Aebersold, R., Burlingame, A. L. & Bradshaw, R. A. 2013. Western blots versus selected reaction monitoring assays: time to turn the tables? *Mol Cell Proteomics,* 12**,** 2381-2.

Aebersold, R. & Mann, M. 2003. Mass spectrometry-based proteomics. *Nature,* 422**,** 198-207.

Aebersold, R. & Mann, M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature,* 537**,** 347-55.

Agard, N. J., Mahrus, S., Trinidad, J. C., Lynn, A., Burlingame, A. L. & Wells, J. A. 2012. Global kinetic analysis of proteolysis via quantitative targeted proteomics. *Proc Natl Acad Sci U S A,* 109**,** 1913-8.

Agard, N. J., Maltby, D. & Wells, J. A. 2010. Inflammatory stimuli regulate caspase substrate profiles. *Mol Cell Proteomics,* 9**,** 880-93.

Agard, N. J. & Wells, J. A. 2009. Methods for the proteomic identification of protease substrates. *Curr Opin Chem Biol,* 13**,** 503-9.

Agarwal, P., van der Weijden, J., Sletten, E. M., Rabuka, D. & Bertozzi, C. R. 2013. A Pictet-Spengler ligation for protein chemical modification. *Proc Natl Acad Sci U S A,* 110**,** 46-51.

Aichler, M. & Walch, A. 2015. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Lab Invest,* 95**,** 422-31.

Altelaar, A. F., Munoz, J. & Heck, A. J. 2013. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet,* 14**,** 35-48.

Andersson, L. & Porath, J. 1986. Isolation of phosphoproteins by immobilized metal (Fe^{3+}) affinity chromatography. *Analytical biochemistry,* 154**,** 250-254.

Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Lo Muzio, L. 2017. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *Int J Mol Med,* 40**,** 271-280.

Arnesen, T. 2011. Towards a functional understanding of protein N-terminal acetylation. *PLoS Biol,* 9**,** e1001074.

Arnesen, T., Van Damme, P., Polevoda, B., Helsens, K., Evjenth, R., Colaert, N., . . . Gevaert, K. 2009. Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. *Proc Natl Acad Sci U S A,* 106**,** 8157-62.

ATCC. 1925. *Jurkat, Clone E6-1 (ATCC® TIB-152™)* [Online]. Manassas (VA): American Type Culture Collection (ATCC). Available: https://www.atcc.org/products/all/TIB-152 [Accessed 12 Jan. 2018].

Bachmann, W. E. & Barton, M. X. 1938. The relative proportions of stereoisomeric oximes formed in the oximation of unsymmetrical ketones. *The Journal of Organic Chemistry,* 03**,** 300-311.

Bairoch, A. & Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res,* 28**,** 45-8.

Baker, D. P., Lin, E. Y., Lin, K., Pellegrini, M., Petter, R. C., Chen, L. L., . . . Pepinsky, R. B. 2006. N-terminally PEGylated human interferon-beta-1a with improved pharmacokinetic properties and *in vivo* efficacy in a melanoma angiogenesis model. *Bioconjug Chem,* 17**,** 179-88.

Bantscheff, M., Boesche, M., Eberhard, D., Matthieson, T., Sweetman, G. & Kuster, B. 2008. Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol Cell Proteomics,* 7**,** 1702-13.

Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. 2012. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem,* 404**,** 939-65.

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

Bass, J. I. F., Pons, C., Kozlowski, L., Reece-Hoyes, J. S., Shrestha, S., Holdorf, A. D., . . . Walhout, A. J. 2016. A gene-centered *C. elegans* protein–DNA interaction network provides a framework for functional predictions. *Molecular systems biology,* 12**,** 884.

Beardsley, R. L., Karty, J. A. & Reilly, J. P. 2000. Enhancing the intensities of lysine-terminated tryptic peptide ions in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Communications in Mass Spectrometry,* 14**,** 2147-2153.

Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., . . . Olsen, J. V. 2017. An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst,* 4**,** 587-599 e4.

Bellac, C. L., Dufour, A., Krisinger, M. J., Loonchanta, A., Starr, A. E., Auf Dem Keller, U., . . . Overall, C. M. 2014. Macrophage matrix metalloproteinase-12 dampens inflammation and neutrophil influx in arthritis. *Cell Rep,* 9**,** 618-32.

Berg, J. M., Tymoczko, J. L. & Stryer, L. 2012. *Biochemistry,* New York, W.H. Freeman.

Biemann, K. 1986. Mass spectrometric methods for protein sequencing. *Anal Chem,* 58**,** 1288A-1300A.

Bilbao, A., Varesio, E., Luban, J., Strambio-De-Castillia, C., Hopfgartner, G., Muller, M. & Lisacek, F. 2015. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics,* 15**,** 964-80.

Bland, C., Bellanger, L. & Armengaud, J. 2014a. Magnetic immunoaffinity enrichment for selective capture and MS/MS analysis of N-terminal-TMPP-labeled peptides. *J Proteome Res,* 13**,** 668-80.

Bland, C., Hartmann, E. M., Christie-Oleza, J. A., Fernandez, B. & Armengaud, J. 2014b. N-terminal-oriented proteogenomics of the marine bacterium *roseobacter denitrificans* Och114 using N-Succinimidyloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP) labeling and diagonal chromatography. *Mol Cell Proteomics,* 13**,** 1369-81.

Boja, E. S. & Fales, H. M. 2001. Overalkylation of a protein digest with iodoacetamide. *Anal Chem,* 73**,** 3576-82.

Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A. & Pevzner, P. A. 2013. N-terminal protein processing: a comparative proteogenomic analysis. *Mol Cell Proteomics,* 12**,** 14-28.

Bradshaw, R. A., Burlingame, A. L., Carr, S. & Aebersold, R. 2006. Reporting protein identification data the next generation of guidelines. ASBMB.

Brower, V. 2001. Proteomics: biology in the post-genomic era. Companies all over the world rush to lead the way in the new post-genomics race. *EMBO Rep,* 2**,** 558-60.

Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinovic, S. M., Cheng, L. Y., Messner, S., . . . Reiter, L. 2015. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics,* 14**,** 1400-10.

Cabiscol, E., Tamarit, J. & Ros, J. 2014. Protein carbonylation: proteomics, specificity and relevance to aging. *Mass Spectrom Rev,* 33**,** 21-48.

Calvete, J. J. 2014. The expanding universe of mass analyzer configurations for biological analysis. *Methods Mol Biol,* 1072**,** 61-81.

Campa, V. M. & Kypta, R. M. 2011. Issues associated with the use of phosphospecific antibodies to localise active and inactive pools of GSK-3 in cells. *Biol Direct,* 6**,** 4.

Canfield, R. E. 1963. The amino acid sequence of egg white lysozyme. *J Biol Chem,* 238**,** 2698-707.

Catherman, A. D., Skinner, O. S. & Kelleher, N. L. 2014. Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun,* 445**,** 683-93.

Chan, A. O., Ho, C. M., Chong, H. C., Leung, Y. C., Huang, J. S., Wong, M. K. & Che, C. M. 2012. Modification of N-terminal α-amino groups of peptides and proteins using ketenes. *Journal of the American Chemical Society,* 134**,** 2589-2598.

Chandramouli, K. & Qian, P. Y. 2009. Proteomics: challenges, techniques and possibilities to overcome biological sample complexity. *Hum Genomics Proteomics,* 2009.

Chelius, D. & Bondarenko, P. V. 2002. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res,* 1**,** 317-23.

Chen, B., Brown, K. A., Lin, Z. & Ge, Y. 2018. Top-down proteomics: ready for prime time? *Anal Chem,* 90**,** 110-127.

Chen, S. J., Wu, X., Wadas, B., Oh, J. H. & Varshavsky, A. 2017. An N-end rule pathway that recognizes proline and destroys gluconeogenic enzymes. *Science,* 355**,** eaal3655.

Ciechanover, A. 2005. N-terminal ubiquitination. *Methods Mol Biol,* 301**,** 255-70.

Cipriani, F., Rower, M., Landret, C., Zander, U., Felisaz, F. & Marquez, J. A. 2012. CrystalDirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. *Acta Crystallogr D Biol Crystallogr,* 68**,** 1393-9.

Coffey, C. M. & Gronert, S. 2016. A cleavable biotin tagging reagent that enables the enrichment and identification of carbonylation sites in proteins. *Anal Bioanal Chem,* 408**,** 865-74.

Cohen, L. 1968. Group-specific reagents in protein chemistry. *Annual review of biochemistry,* 37**,** 695-726.

Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. 2009. Improved visualization of protein consensus sequences by iceLogo. *Nat Methods,* 6**,** 786-7.

Cooks, R. G., Glish, G. L., McLuckey, S. A. & Kaiser, R. E. 1991. Ion trap mass spectrometry. *Chemical & Engineering News Archive,* 69**,** 26-41.

Cox, J. & Mann, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol,* 26**,** 1367-72.

Creese, A. J. & Cooper, H. J. 2007. Liquid chromatography electron capture dissociation tandem mass spectrometry (LC-ECD-MS/MS) versus liquid chromatography collision-induced dissociation tandem mass spectrometry (LC-CID-MS/MS) for the identification of proteins. *J Am Soc Mass Spectrom,* 18**,** 891-7.

Crick, F. 1970. Central dogma of molecular biology. *Nature,* 227**,** 561-3.

Cuatrecasas, P., Wilchek, M. & Anfinsen, C. B. 1968. Selective enzyme purification by affinity chromatography. *Proc Natl Acad Sci U S A,* 61**,** 636-43.

Dalle-Donne, I., Rossi, R., Giustarini, D., Milzani, A. & Colombo, R. 2003. Protein carbonyl groups as biomarkers of oxidative stress. *Clin Chim Acta,* 329**,** 23-38.

Davidson, G. R., Armstrong, S. D. & Beynon, R. J. 2011. Positional proteomics at the N-terminus as a means of proteome simplification. *Gel-Free Proteomics.* Springer.

Dawson, P. E., Muir, T. W., Clark-Lewis, I. & Kent, S. B. 1994. Synthesis of proteins by native chemical ligation. *Science,* 266**,** 776-9.

Dawson, P. H. 2013. *Quadrupole mass spectrometry and its applications*, Elsevier.

Dawson, R. M. C., Elliott, D. C., Elliott, W. H. & Jones, K. M. 2002. *Data for biochemical research*, Clarendon Press.

de Graaf, E. L., Altelaar, A. F., van Breukelen, B., Mohammed, S. & Heck, A. J. 2011. Improving SRM assay development: a global comparison between triple quadrupole, ion trap, and higher energy CID peptide fragmentation spectra. *J Proteome Res,* 10**,** 4334-41.

de Hoffmann, E. & Stroobant, V. 2007. Mass spectrometry: principles and applications. John Wiley & Sons, Inc.

Demir, F., Niedermaier, S., Kizhakkedathu, J. N. & Huesgen, P. F. 2017. Profiling of protein N-termini and their modifications in complex samples. *Methods Mol Biol,* 1574**,** 35-50.

Deng, J., Zhang, G., Huang, F. K. & Neubert, T. A. 2015. Identification of protein N-termini using TMPP or dimethyl labeling and mass spectrometry. *Methods Mol Biol,* 1295**,** 249-58.

Denisov, E., Damoc, E., Lange, O. & Makarov, A. 2012. Orbitrap mass spectrometry with resolving powers above 1,000,000. *International Journal of Mass Spectrometry,* 325-327**,** 80-85.

DePhillips, P., Lagerlund, I., Farenmark, J. & Lenhoff, A. M. 2004. Effect of spacer arm length on protein retention on a strong cation exchange adsorbent. *Anal Chem,* 76**,** 5816-22.

Dix, M. M., Simon, G. M. & Cravatt, B. F. 2008. Global mapping of the topography and magnitude of proteolytic events in apoptosis. *Cell,* 134**,** 679-91.

Dixon, H. B. 1964. Transamination of peptides *Biochemical Journal,* 92**,** 661-6.

Dixon, H. B. 1984. N-terminal modification of proteins-a review. *Journal of Protein Chemistry,* 3**,** 99-108.

Dixon, H. B. & Fields, R. 1972. Specific modification of NH_2-terminal residues by transamination. *Methods in Enzymology,* 25**,** 409-419.

Dixon, H. B. & Moret, V. 1964. Removal of the N-terminal residue of corticotrophin. *Biochemical Journal,* 93**,** 25C-26C.

Dixon, H. B. & Moret, V. 1965. Removal of the N-terminal residue of a protein after transamination. *The Biochemical journal,* 94**,** 463-469.

Dixon, H. B. & Perham, R. N. 1968. Reversible blocking of amino groups with citraconic anhydride. *Biochem J,* 109**,** 312-4.

Dixon, H. B. & Weitkamp, L. R. 1962. Conversion of the N-terminal serine residue of corticotrophin into glycine. *Biochem J,* 84**,** 462-8.

Doerr, A. 2014. DIA mass spectrometry. *Nat Methods,* 12**,** 35.

Dole, M., Mack, L. L., Hines, R. L., Mobley, R. C., Ferguson, L. D. & Alice, M. B. 1968. Molecular beams of macroions. *The Journal of Chemical Physics,* 49**,** 2240-2249.

Domon, B. & Aebersold, R. 2006. Mass spectrometry and protein analysis. *Science,* 312**,** 212-7.

Drazic, A., Myklebust, L. M., Ree, R. & Arnesen, T. 2016. The world of protein acetylation. *Biochim Biophys Acta,* 1864**,** 1372-401.

Dubrovska, A. 2009. Efficient enrichment of intact phosphorylated proteins by modified immobilized metal-affinity chromatography. *The Protein Protocols Handbook.* Springer.

Dundas, C. M., Demonte, D. & Park, S. 2013. Streptavidin-biotin technology: improvements and innovations in chemical and biological applications. *Appl Microbiol Biotechnol,* 97**,** 9343-53.

Dunham, W. H., Mullin, M. & Gingras, A. C. 2012. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics,* 12**,** 1576-90.

Eckhard, U., Marino, G., Butler, G. S. & Overall, C. M. 2016. Positional proteomics in the era of the human proteome project on the doorstep of precision medicine. *Biochimie,* 122**,** 110-8.

Edge, A. S. 2003. Deglycosylation of glycoproteins with trifluoromethanesulphonic acid: elucidation of molecular structure and function. *Biochemical Journal,* 376**,** 339.

El-Aneed, A., Cohen, A. & Banoub, J. 2009. Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers. *Applied Spectroscopy Reviews,* 44**,** 210-230.

Fedorova, M., Bollineni, R. C. & Hoffmann, R. 2014. Protein carbonylation as a major hallmark of oxidative damage: update of analytical strategies. *Mass Spectrom Rev,* 33**,** 79-97.

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

Fields, R. & Dixon, H. B. 1968. A spectrophotometric method for the microdetermination of periodate. *The Biochemical journal,* 108**,** 883-887.

Flavin, W. P., Bousset, L., Green, Z. C., Chu, Y., Skarpathiotis, S., Chaney, M. J., . . . Campbell, E. M. 2017. Endocytic vesicle rupture is a conserved mechanism of cellular invasion by amyloid proteins. *Acta Neuropathol,* 134**,** 629-653.

Gawron, D., Ndah, E., Gevaert, K. & Van Damme, P. 2016. Positional proteomics reveals differences in N-terminal proteoform stability. *Mol Syst Biol,* 12**,** 858.

Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. 2012. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics,* 11**,** M111 014050.

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

Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R. & Vandekerckhove, J. 2003. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol,* 21**,** 566-9.

Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R., Hoorelbeke, B., Demol, H., . . . Vandekerckhove, J. 2002. Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis identification of more than 800 Escherichia coli proteins. *Molecular & cellular proteomics,* 1**,** 896-903.

Giglione, C., Fieulaine, S. & Meinnel, T. 2015. N-terminal protein modifications: Bringing back into play the ribosome. *Biochimie,* 114**,** 134-46.

Gill, G. 2004. SUMO and ubiquitin in the nucleus: different functions, similar mechanisms? *Genes & development,* 18**,** 2046-2059.

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., . . . Aebersold, R. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics,* 11**,** O111 016717.

Gillis, S. & Watson, J. 1980. Biochemical and biological characterization of lymphocyte regulatory molecules. V. Identification of an interleukin 2-producing human leukemia T cell line. *J Exp Med,* 152**,** 1709-19.

Gilmore, J. M., Scheck, R. A., Esser-Kahn, A. P., Joshi, N. S. & Francis, M. B. 2006. N-terminal protein modification through a biomimetic transamination reaction. *Angewandte Chemie International Edition,* 45**,** 5307-5311.

Gonzalez, M., Argarana, C. E. & Fidelio, G. D. 1999. Extremely high thermal stability of streptavidin and avidin upon biotin binding. *Biomol Eng,* 16**,** 67-72.

Green, N. M. 1975. Avidin. *Adv Protein Chem,* 29**,** 85-133.

Griffiths, J. 2008. A brief history of mass spectrometry. *Anal Chem,* 80**,** 5678-83.

Grimsley, G. R., Scholtz, J. M. & Pace, C. N. 2009. A summary of the measured p*K* values of the ionizable groups in folded proteins. *Protein Sci,* 18**,** 247-51.

Gross, J. H. 2017. *Mass spectrometry: a textbook*, Springer.

Gupta, N. & Pevzner, P. A. 2009. False discovery rates of protein identifications: a strike against the two-peptide rule. *J Proteome Res,* 8**,** 4173-81.

Guthals, A. & Bandeira, N. 2012. Peptide identification by tandem mass spectrometry with alternate fragmentation modes. *Mol Cell Proteomics,* 11**,** 550-7.

Hart-Smith, G. & Blanksby, S. J. 2012. Mass analysis. *Mass spectrometry in polymer chemistry***,** 5-32.

Hartmann, E. M. & Armengaud, J. 2014. N-terminomics and proteogenomics, getting off to a good start. *Proteomics,* 14**,** 2637-46.

Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S. & Coon, J. J. 2014. The one hour yeast proteome. *Mol Cell Proteomics,* 13**,** 339-47.

Helsens, K., Van Damme, P., Degroeve, S., Martens, L., Arnesen, T., Vandekerckhove, J. & Gevaert, K. 2011. Bioinformatics analysis of a *Saccharomyces cerevisiae* N-terminal proteome provides evidence of alternative translation initiation and post-translational N-terminal acetylation. *J Proteome Res,* 10**,** 3578-89.

Hendriks, I. A. & Vertegaal, A. C. 2016. A comprehensive compilation of SUMO proteomics. *Nat Rev Mol Cell Biol,* 17**,** 581-95.

Henrikson, K. P., Allen, S. H. & Maloy, W. L. 1979. An avidin monomer affinity column for the purification of biotin-containing enzymes. *Anal Biochem,* 94**,** 366-70.

Hermanson, G. 1996. *Bioconjugate techniques,* San Diego, CA, Academic Press.

Holding, A. N. 2015. XL-MS: protein cross-linking coupled with mass spectrometry. *Methods,* 89**,** 54-63.

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

Huang, D. W., Sherman, B. T. & Lempicki, R. A. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols,* 4**,** 44.

Huang, T., Wang, J., Yu, W. & He, Z. 2012. Protein inference: a review. *Briefings in Bioinformatics,* 13**,** 586-614.

Huang, Z. H., Wu, J., Roth, K. D., Yang, Y., Gage, D. A. & Watson, J. T. 1997. A picomole-scale method for charge derivatization of peptides for sequence analysis by mass spectrometry. *Anal Chem,* 69**,** 137-44.

Hughes, G. J., Frutiger, S., Paquet, N., Pasquali, C., Sanchez, J. C., Tissot, J. D., . . . Hochstrasser, D. F. 1993. Human liver protein map: update 1993. *Electrophoresis,* 14**,** 1216-22.

Hunt, D. F., Yates, J. R., 3rd, Shabanowitz, J., Winston, S. & Hauer, C. R. 1986. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci U S A,* 83**,** 6233-7.

Huse, M. 2009. The T-cell-receptor signaling network. *J Cell Sci,* 122**,** 1269-73.

Impens, F., Colaert, N., Helsens, K., Plasman, K., Van Damme, P., Vandekerckhove, J. & Gevaert, K. 2010. MS-driven protease substrate degradomics. *Proteomics,* 10**,** 1284-96.

Impens, F., Radoshevich, L., Cossart, P. & Ribet, D. 2014. Mapping of SUMO sites and analysis of SUMOylation changes induced by external stimuli. *Proceedings of the National Academy of Sciences of the United States of America,* 111**,** 12432-12437.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. & Mann, M. 2005. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics,* 4**,** 1265-72.

Jarnuczak, A. F., Lee, D. C., Lawless, C., Holman, S. W., Eyers, C. E. & Hubbard, S. J. 2016. Analysis of intrinsic peptide detectability via integrated label-free and SRM-based absolute quantitative proteomics. *J Proteome Res,* 15**,** 2945-59.

Johnson, J. V., Yost, R. A., Kelley, P. E. & Bradford, D. C. 1990. Tandem-in-space and tandem-in-time mass spectrometry: triple quadrupoles and quadrupole ion traps. *Analytical Chemistry,* 62**,** 2162-2172.

Johnson, R. S. & Biemann, K. 1989. Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed Environ Mass Spectrom,* 18**,** 945-57.

Kalia, J. & Raines, R. T. 2008. Hydrolytic stability of hydrazones and oximes. *Angew Chem Int Ed Engl,* 47**,** 7523-6.

Kalvik, T. V. & Arnesen, T. 2013. Protein N-terminal acetyltransferases in cancer. *Oncogene,* 32**,** 269-76.

Kang, D., Nam, H., Kim, Y. S. & Moon, M. H. 2005. Dual-purpose sample trap for on-line strong cation-exchange chromatography/reversed-phase liquid chromatography/tandem mass spectrometry for shotgun proteomics. Application to the human Jurkat T-cell proteome. *J Chromatogr A,* 1070**,** 193-200.

Karas, M., Gluckmann, M. & Schafer, J. 2000. Ionization in matrix-assisted laser desorption/ionization: singly charged molecular ions are the lucky survivors. *J Mass Spectrom,* 35**,** 1-12.

Karas, M. & Hillenkamp, F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem,* 60**,** 2299-301.

Kim, J. S., Dai, Z., Aryal, U. K., Moore, R. J., Camp, D. G., 2nd, Baker, S. E., . . . Qian, W. J. 2013. Resin-assisted enrichment of N-terminal peptides for characterizing proteolytic processing. *Anal Chem,* 85**,** 6826-32.

Kim, M., Eetemadi, A. & Tagkopoulos, I. 2017. DeepPep: Deep proteome inference from peptide profiles. *PLoS computational biology,* 13**,** e1005661.

Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., . . . Pandey, A. 2014. A draft map of the human proteome. *Nature,* 509**,** 575-81.

Kleifeld, O., Doucet, A., Auf Dem Keller, U., Prudova, A., Schilling, O., Kainthan, R. K., . . . Overall, C. M. 2010. Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat Biotechnol,* 28**,** 281-8.

Klein, T., Fung, S. Y., Renner, F., Blank, M. A., Dufour, A., Kang, S., . . . Overall, C. M. 2015. The paracaspase MALT1 cleaves HOIL1 reducing linear ubiquitination by LUBAC to dampen lymphocyte NF-kappaB signalling. *Nat Commun,* 6**,** 8777.

Krusemark, C. J., Frey, B. L., Belshaw, P. J. & Smith, L. M. 2009. Modifying the charge state distribution of proteins in electrospray ionization mass spectrometry by chemical derivatization. *J Am Soc Mass Spectrom,* 20**,** 1617-25.

Kuroishi, T., Rios-Avila, L., Pestinger, V., Wijeratne, S. S. & Zempleni, J. 2011. Biotinylation is a natural, albeit rare, modification of human histones. *Mol Genet Metab,* 104**,** 537-45.

Kyoto Encyclopedia of Genes and Genomes. 1995. *KEGG: Kyoto Encyclopedia of Genes and Genomes.* [Online]. Available: http://www.kegg.jp [Accessed 12 January 2018].

Kyte, J. & Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol,* 157**,** 105-32.

Lai, Z. W., Gomez-Auli, A., Keller, E. J., Mayer, B., Biniossek, M. L. & Schilling, O. 2015. Enrichment of protein N-termini by charge reversal of internal peptides. *Proteomics,* 15**,** 2470-8.

Lange, O., Damoc, E., Wieghaus, A. & Makarov, A. 2014a. Enhanced Fourier transform for Orbitrap mass spectrometry. *International Journal of Mass Spectrometry,* 369**,** 16-22.

Lange, P. F., Huesgen, P. F., Nguyen, K. & Overall, C. M. 2014b. Annotating N termini for the human proteome project: N termini and Nalpha-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. *J Proteome Res,* 13**,** 2028-44.

Lange, P. F. & Overall, C. M. 2013. Protein TAILS: when termini tell tales of proteolysis and function. *Curr Opin Chem Biol,* 17**,** 73-82.

Lange, V., Picotti, P., Domon, B. & Aebersold, R. 2008. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol,* 4**,** 222.

Lee, K. E., Heo, J. E., Kim, J. M. & Hwang, C. S. 2016. N-terminal acetylation-targeted N-end rule proteolytic system: the Ac/N-end rule pathway. *Mol Cells,* 39**,** 169-78.

Li, Y. F., Arnold, R. J., Tang, H. & Radivojac, P. 2010. The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *J Proteome Res,* 9**,** 6288-97.

Lim, W. K., Rosgen, J. & Englander, S. W. 2009. Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. *Proc Natl Acad Sci U S A,* 106**,** 2595-600.

Liongue, C., John, L. B. & Ward, A. 2011. Origins of adaptive immunity. *Crit Rev Immunol,* 31**,** 61-71.

Liu, H., Zhang, J., Sun, H., Xu, C., Zhu, Y. & Xie, H. 2011. The prediction of peptide charge states for electrospray ionization in mass spectrometry. *Procedia Environmental Sciences,* 8**,** 483-491.

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

Luo, S. & Wehr, N. B. 2009. Protein carbonylation: avoiding pitfalls in the 2,4-dinitrophenylhydrazine assay. *Redox Rep,* 14**,** 159-66.

MacDonald, J. I., Munch, H. K., Moore, T. & Francis, M. B. 2015. One-step site-specific modification of native proteins with 2-pyridinecarboxyaldehydes. *Nat Chem Biol,* 11**,** 326-31.

Macek, B., Mann, M. & Olsen, J. V. 2009. Global and site-specific quantitative phosphoproteomics: Principles and applications. *Annual Review of Pharmacology and Toxicology.*

MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., . . . MacCoss, M. J. 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics,* 26**,** 966-8.

Maes, K., Smolders, I., Michotte, Y. & Van Eeckhaut, A. 2014. Strategies to reduce aspecific adsorption of peptides and proteins in liquid chromatography-mass spectrometry based bioanalyses: an overview. *J Chromatogr A,* 1358**,** 1-13.

Mahrus, S., Trinidad, J. C., Barkan, D. T., Sali, A., Burlingame, A. L. & Wells, J. A. 2008. Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini. *Cell,* 134**,** 866-76.

Makhatadze, G. I. & Privalov, P. L. 1992. Protein interactions with urea and guanidinium chloride: a calorimetric study. *Journal of molecular biology,* 226**,** 491-505.

Maleknia, S. D. & Johnson, R. 2011. Mass spectrometry of amino acids and proteins. *Amino Acids, Peptides and Proteins in Organic Chemistry.* Wiley-VCH Verlag GmbH & Co. KGaA.

Maley, F., Trimble, R. B., Tarentino, A. L. & Plummer Jr, T. H. 1989. Characterization of glycoproteins and their associated oligosaccharides through the use of endoglycosidases. *Analytical biochemistry,* 180**,** 195-204.

Martinez, A., Traverso, J. A., Valot, B., Ferro, M., Espagne, C., Ephritikhine, G., . . . Meinnel, T. 2008. Extent of N-terminal modifications in cytosolic proteins from eukaryotes. *Proteomics,* 8**,** 2809-31.

Marttila, A. T., Laitinen, O. H., Airenne, K. J., Kulik, T., Bayer, E. A., Wilchek, M. & Kulomaa, M. S. 2000. Recombinant NeutraLite avidin: a non-glycosylated, acidic mutant of chicken avidin that exhibits high affinity for biotin and low non-specific binding properties. *FEBS Lett,* 467**,** 31-6.

Matlin, S. A., Jiang, L. X., Roshdy, S. & Zhou, R. H. 1990. Resolution and identification of steroid oxime syn and antiisomers by HPLC. *Journal of Liquid Chromatography,* 13**,** 3455-3463.

Matthews, D. J. & Wells, J. A. 1993. Substrate phage: selection of protease substrates by monovalent phage display. *Science,* 260**,** 1113-7.

McDonald, L. & Beynon, R. J. 2006. Positional proteomics: preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nat Protoc,* 1**,** 1790-8.

McDonald, L., Robertson, D. H., Hurst, J. L. & Beynon, R. J. 2005. Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat Methods,* 2**,** 955-7.

McLuckey, S. A. & Huang, T. Y. 2009. Ion/ion reactions: new chemistry for analytical MS. *Anal Chem,* 81**,** 8669-76.

Mejía-Manzano, L. A., González-Valdez, J., Mayolo-Deloisa, K., Escalante-Vázquez, E. J. & Rito-Palomares, M. 2016. Covalent immobilization of antibodies for the preparation of immunoaffinity chromatographic supports. *Separation Science and Technology,* 51**,** 1736-1743.

Michalski, A., Cox, J. & Mann, M. 2011. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res,* 10**,** 1785-93.

Mikesh, L. M., Ueberheide, B., Chi, A., Coon, J. J., Syka, J. E., Shabanowitz, J. & Hunt, D. F. 2006. The utility of ETD mass spectrometry in proteomic analysis. *Biochim Biophys Acta,* 1764**,** 1811-22.

Mirzaei, H. & Regnier, F. 2006. Enhancing electrospray ionization efficiency of peptides by derivatization. *Anal Chem,* 78**,** 4175-83.

Mix, H. & Wilcke, F. W. 1960. Nichtenzymatische reaktionen zwischen α-amino- und α-ketosäuren, II1: durch Kupfer(II)-Ionen und pyridin katalysierte umsetzungen zwischen α-aminosäuren und pyruvat. *Hoppe-Seyler's Zeitschrift fur Physiologische Chemie,* 318**,** 148-158.

Mommen, G. P., van de Waterbeemd, B., Meiring, H. D., Kersten, G., Heck, A. J. & de Jong, A. P. 2012. Unbiased selective isolation of protein N-terminal peptides from complex proteome samples using phospho tagging (PTAG) and TiO(2)-based depletion. *Mol Cell Proteomics,* 11**,** 832-42.

Hensman Moss, D. J., Flower, M. D., Lo, K. K., Miller, J. R., van Ommen, G. J., 't Hoen, P. A., . . . Tabrizi, S. J. 2017. Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. *Scientific reports,* 7**,** 44849.

Murn, J. & Shi, Y. 2017. The winding path of protein methylation research: milestones and new frontiers. *Nat Rev Mol Cell Biol,* 18**,** 517-527.

Mustelin, T. & Tasken, K. 2003. Positive and negative regulation of T-cell activation through kinases and phosphatases. *Biochem J,* 371**,** 15-27.

Myers, J. K., Pace, C. N. & Scholtz, J. M. 1995. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci,* 4**,** 2138-48.

Nguyen, T. D., Carrascal, M., Vidal-Cortes, O., Gallardo, O., Casas, V., Gay, M., . . . Abian, J. 2016. The phosphoproteome of human Jurkat T cell clones upon costimulation with anti-CD3/anti-CD28 antibodies. *J Proteomics,* 131**,** 190-198.

Ning, X., Temming, R. P., Dommerholt, J., Guo, J., Ania, D. B., Debets, M. F., . . . van Delft, F. L. 2010. Protein modification by strain-promoted alkyne-nitrone cycloaddition. *Angew Chem Int Ed Engl,* 49**,** 3065-8.

Obermeyer, A. C., Jarman, J. B. & Francis, M. B. 2014. N-terminal modification of proteins with o-aminophenols. *J Am Chem Soc,* 136**,** 9572-9.

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

Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. & Mann, M. 2006. Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell,* 127**,** 635-48.

Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S. & Mann, M. 2007. Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods,* 4**,** 709-12.

Olsen, J. V. & Mann, M. 2013. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics,* 12**,** 3444-52.

Olsen, J. V., Ong, S. E. & Mann, M. 2004. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics,* 3**,** 608-14.

Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. & 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**,** 376-86.

Ong, S. E. & Mann, M. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol,* 1**,** 252-62.

Ordureau, A., Munch, C. & Harper, J. W. 2015. Quantifying ubiquitin signaling. *Mol Cell,* 58**,** 660-76.

Oregioni, A., Stieglitz, B., Kelly, G., Rittinger, K. & Frenkiel, T. 2017. Determination of the pK_a of the N-terminal amino group of ubiquitin by NMR. *Sci Rep,* 7**,** 43748.

Osaka, I. & Takayama, M. 2014. Influence of hydrophobicity on positive- and negative-ion yields of peptides in electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom,* 28**,** 2222-6.

Osorio, D., Rondón-Villarrea, P. & Torres, R. 2015. Peptides: a package for data mining of antimicrobial peptides. *R Journal,* 7**,** 4-14.

Papanikos, A., Rademann, J. & Meldal, M. 2001. alpha-Ketocarbonyl peptides: a general approach to reactive resin-bound intermediates in the synthesis of peptide isosteres for protease inhibitor screening on solid support. *J Am Chem Soc,* 123**,** 2176-81.

Pawson, T. & Scott, J. D. 2005. Protein phosphorylation in signaling--50 years and counting. *Trends Biochem Sci,* 30**,** 286-90.

Peng, J. & Gygi, S. P. 2001. Proteomics: the move to mixtures. *J Mass Spectrom,* 36**,** 1083-91.

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

Picotti, P. & Aebersold, R. 2012. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods,* 9**,** 555-66.

Plasman, K., Demol, H., Bird, P. I., Gevaert, K. & Van Damme, P. 2014. Substrate specificities of the granzyme tryptases A and K. *Journal of Proteome Research,* 13**,** 6067-6077.

Plasman, K., Van Damme, P. & Gevaert, K. 2013. Contemporary positional proteomics strategies to study protein processing. *Curr Opin Chem Biol,* 17**,** 66-72.

Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M. & Nicholson, J. K. 2006. UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Commun Mass Spectrom,* 20**,** 1989-94.

Prudova, A., Auf Dem Keller, U., Butler, G. S. & Overall, C. M. 2010. Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. *Mol Cell Proteomics,* 9**,** 894-911.

Pubmed. 1996. *PubMed* [Online]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. Available: https://www.ncbi.nlm.nih.gov/pubmed/ [Accessed 12 Jan. 2018].

Purvine, S., Eppel, J. T., Yi, E. C. & Goodlett, D. R. 2003. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics,* 3**,** 847-50.

Qi, Y. & Volmer, D. A. 2016. Structural analysis of small to medium-sized molecules by mass spectrometry after electron-ion fragmentation (ExD) reactions. *Analyst,* 141**,** 794-806.

Qian, W. J., Jacobs, J. M., Liu, T., Camp, D. G., 2nd & Smith, R. D. 2006. Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Mol Cell Proteomics,* 5**,** 1727-44.

Rees, J. S., Lilley, K. S. & Jackson, A. P. 2015. SILAC-iPAC: a quantitative method for distinguishing genuine from non-specific components of protein complexes by parallel affinity capture. *J Proteomics,* 115**,** 143-56.

Rees, J. S., Lowe, N., Armean, I. M., Roote, J., Johnson, G., Drummond, E., . . . Lilley, K. S. 2011. *In vivo* analysis of proteomes and interactomes using Parallel Affinity Capture (iPAC) coupled to mass spectrometry. *Mol Cell Proteomics,* 10**,** M110 002386.

Reid, E. 1951. Potentiation by adrenocorticotrophin of the diabetogenic action of growth-hormone preparations. *Nature,* 168**,** 878.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Seraphin, B. 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol,* 17**,** 1030-2.

Rock, K. L., Gramm, C., Rothstein, L., Clark, K., Stein, R., Dick, L., . . . Goldberg, A. L. 1994. Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell,* 78**,** 761-71.

Roepstorff, P. & Fohlman, J. 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom,* 11**,** 601.

Rogers, L. D. & Overall, C. M. 2013. Proteolytic post-translational modification of proteins: proteomic tools and methodology. *Mol Cell Proteomics,* 12**,** 3532-42.

Rogowska-Wrzesinska, A., Wojdyla, K., Nedic, O., Baron, C. P. & Griffiths, H. R. 2014. Analysis of protein carbonylation--pitfalls and promise in commonly used methods. *Free Radic Res,* 48**,** 1145-62.

Rosen, C. B. & Francis, M. B. 2017. Targeting the N terminus for site-selective protein modification. *Nature Chemical Biology,* 13**,** 697-705.

Ross, C. A. & Poirier, M. A. 2004. Protein aggregation and neurodegenerative disease. *Nat Med,* 10 Suppl**,** S10-7.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., . . . Pappin, D. J. 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics,* 3**,** 1154-69.

Rost, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S. M., Schubert, O. T., . . . Aebersold, R. 2014. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol,* 32**,** 219-23.

Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., . . . Comb, M. J. 2005. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature biotechnology,* 23**,** 94.

Said, H. M. 2012. Biotin: biochemical, physiological and clinical aspects. *Subcell Biochem,* 56**,** 1-19.

Sanman, L. E. & Bogyo, M. 2014. Activity-based profiling of proteases. *Annu Rev Biochem,* 83**,** 249-73.

Sarkar, B., Dixon, H. B. & Webster, D. 1978. Removal by transamination and scission of residues from the peptide representing the copper-transport site of serum albumin. *Biochemical Journal,* 173**,** 895-897.

Sasaki, T., Kodama, K., Suzuki, H., Fukuzawa, S. & Tachibana, K. 2008. N-terminal labeling of proteins by the Pictet-Spengler reaction. *Bioorg Med Chem Lett,* 18**,** 4550-3.

Schlage, P., Egli, F. E., Nanni, P., Wang, L. W., Kizhakkedathu, J. N., Apte, S. S. & auf dem Keller, U. 2014. Time-resolved analysis of the matrix metalloproteinase 10 substrate degradome. *Mol Cell Proteomics,* 13**,** 580-93.

Schneider, U., Schwenk, H. U. & Bornkamm, G. 1977. Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *Int J Cancer,* 19**,** 621-6.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., . . . Selbach, M. 2011. Global quantification of mammalian gene expression control. *Nature,* 473**,** 337-42.

Selo, I., Negroni, L., Creminon, C., Grassi, J. & Wal, J. M. 1996. Preferential labeling of alpha-amino N-terminal groups in peptides by biotin: application to the detection of specific anti-peptide antibodies by enzyme immunoassays. *J Immunol Methods,* 199**,** 127-38.

Sharma, K., D'souza, R. C., Tyanova, S., Schaab, C., Wisniewski, J. R., Cox, J. & Mann, M. 2014. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep,* 8**,** 1583-94.

Shema, G., Nguyen, M. T., Solari, F. A., Loroch, S., Venne, A. S., Kollipara, L., . . . Zahedi, R. P. 2018. Simple, scalable and ultra-sensitive tip-based identification of protease substrates. *Molecular & Cellular Proteomics***,** mcp. TIR117. 000302.

Shi, Y., van Nostrum, C. F. & Hennink, W. E. 2015. Interfacially hydrazone cross-linked thermosensitive polymeric micelles for acid-triggered release of paclitaxel. *ACS Biomaterials Science & Engineering,* 1**,** 393-404.

Slechtova, T., Gilar, M., Kalikova, K. & Tesarova, E. 2015. Insight into trypsin miscleavage: comparison of kinetic constants of problematic peptide sequences. *Anal Chem,* 87**,** 7636-43.

Sleno, L. & Volmer, D. A. 2004. Ion activation methods for tandem mass spectrometry. *J Mass Spectrom,* 39**,** 1091-112.

Sonomura, K., Kuyama, H., Matsuo, E., Tsunasawa, S., Futaki, S. & Nishimura, O. 2011. Selective isolation of N-blocked peptide by combining AspN digestion, transamination, and tosylhydrazine glass treatment. *Anal Biochem,* 410**,** 214-23.

Sonomura, K., Kuyama, H., Matsuo, E., Tsunasawa, S. & Nishimura, O. 2009a. A method for terminus proteomics: selective isolation and labeling of N-terminal peptide from protein through transamination reaction. *Bioorg Med Chem Lett,* 19**,** 6544-7.

Sonomura, K., Kuyama, H., Matsuo, E., Tsunasawa, S. & Nishimura, O. 2009b. The specific isolation of C-terminal peptides of proteins through a transamination reaction and its advantage for introducing functional groups into the peptide. *Rapid Commun Mass Spectrom,* 23**,** 611-8.

Staes, A., Impens, F., Van Damme, P., Ruttens, B., Goethals, M., Demol, H., . . . Gevaert, K. 2011. Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat Protoc,* 6**,** 1130-41.

Staes, A., Van Damme, P., Timmerman, E., Ruttens, B., Stes, E., Gevaert, K. & Impens, F. 2017. Protease substrate profiling by N-terminal COFRADIC. *Methods Mol Biol,* 1574**,** 51-76.

Steen, H. & Mann, M. 2004. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology,* 5**,** 699.

Sun, S., Zhou, J. Y., Yang, W. & Zhang, H. 2014. Inhibition of protein carbamylation in urea solution using ammonium-containing buffers. *Anal Biochem,* 446**,** 76-81.

Sur, S., Qiao, Y., Fries, A., O'Meally, R. N., Cole, R. N., Kinzler, K. W., . . . Zhou, S. 2015. PRINT: a protein bioconjugation method with exquisite N-terminal specificity. *Sci Rep,* 5**,** 18363.

Swaney, D. L., Wenger, C. D. & Coon, J. J. 2010. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res,* 9**,** 1323-9.

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

Tanco, S., Gevaert, K. & Van Damme, P. 2015. C-terminomics: Targeted analysis of natural and posttranslationally modified protein and peptide C-termini. *Proteomics,* 15**,** 903-14.

Taylor, P. J. 2005. Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography-electrospray-tandem mass spectrometry. *Clin Biochem,* 38**,** 328-34.

The Uniprot Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research,* 41**,** D43-D47.

Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., . . . Hamon, C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem,* 75**,** 1895-904.

Thomson, B. & Iribarne, J. 1979. Field induced ion evaporation from liquid surfaces at atmospheric pressure. *The Journal of Chemical Physics,* 71**,** 4451-4463.

Timmer, J. C. & Salvesen, G. S. 2011. N-terminomics: a high-content screen for protease substrates and their cleavage sites. *Methods Mol Biol,* 753**,** 243-55.

Toney, M. D. 2005. Reaction specificity in pyridoxal phosphate enzymes. *Archives of Biochemistry and Biophysics,* 433**,** 279-287.

Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C. & Nesvizhskii, A. I. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods,* 12**,** 258-64, 7 p following 264.

Tsou, C. C., Tsai, C. F., Teo, G. C., Chen, Y. J. & Nesvizhskii, A. I. 2016. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics,* 16**,** 2257-71.

Turk, B., Turk, D. & Turk, V. 2012. Protease signalling: the cutting edge. *EMBO J,* 31**,** 1630-43.

Tutturen, A. E., Fleckenstein, B. & de Souza, G. A. 2014. Assessing the citrullinome in rheumatoid arthritis synovial fluid with and without enrichment of citrullinated peptides. *J Proteome Res,* 13**,** 2867-73.

Tutturen, A. E., Holm, A. & Fleckenstein, B. 2013. Specific biotinylation and sensitive enrichment of citrullinated peptides. *Anal Bioanal Chem,* 405**,** 9321-31.

Tytgat, H. L., Schoofs, G., Driesen, M., Proost, P., Van Damme, E. J., Vanderleyden, J. & Lebeer, S. 2015. Endogenous biotin-binding proteins: an overlooked factor causing false positives in streptavidin-based protein detection. *Microb Biotechnol,* 8**,** 164-8.

Udeshi, N. D., Pedram, K., Svinkina, T., Fereshetian, S., Myers, S. A., Aygun, O., . . . Carr, S. A. 2017. Antibodies to biotin enable large-scale detection of biotinylation sites on proteins. *Nat Methods,* 14**,** 1167-1170.

Udeshi, N. D., Svinkina, T., Mertins, P., Kuhn, E., Mani, D. R., Qiao, J. W. & Carr, S. A. 2013. Refined preparation and use of anti-diglycine remnant (K-epsilon-GG) antibody enables routine quantification of 10,000s of ubiquitination sites in single proteomics experiments. *Mol Cell Proteomics,* 12**,** 825-31.

Vaca Jacome, A. S., Rabilloud, T., Schaeffer-Reiss, C., Rompais, M., Ayoub, D., Lane, L., . . . Carapito, C. 2015. N-terminome analysis of the human mitochondrial proteome. *Proteomics,* 15**,** 2519-24.

Van Damme, P., Gawron, D., Van Criekinge, W. & Menschaert, G. 2014. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Mol Cell Proteomics,* 13**,** 1245-61.

Van Damme, P., Lasa, M., Polevoda, B., Gazquez, C., Elosegui-Artola, A., Kim, D. S., . . . Aldabe, R. 2012. N-terminal acetylome analyses and functional insights of the N-terminal acetyltransferase NatB. *Proc Natl Acad Sci U S A,* 109**,** 12449-54.

Van Damme, P., Martens, L., Van Damme, J., Hugelier, K., Staes, A., Vandekerckhove, J. & Gevaert, K. 2005. Caspase-specific and nonspecific *in vivo* protein processing during Fas-induced apoptosis. *Nat Methods,* 2**,** 771-7.

van Holde, K. E. 1989. The proteins of chromatin. I. Histones. *Chromatin.* Springer.

Varland, S., Osberg, C. & Arnesen, T. 2015. N-terminal modifications of cellular proteins: The enzymes involved, their substrate specificities and biological effects. *Proteomics,* 15**,** 2385-401.

Varshavsky, A. 1997. The N-end rule pathway of protein degradation. *Genes to cells,* 2**,** 13-28.

Varshavsky, A. 2011. The N-end rule pathway and regulation by proteolysis. *Protein Sci,* 20**,** 1298-345.

Veenstra, T. D., Conrads, T. P. & Issaq, H. J. 2004. What to do with “one-hit wonders”? *Electrophoresis,* 25**,** 1278-1279.

Venne, A. S., Solari, F. A., Faden, F., Paretti, T., Dissmeyer, N. & Zahedi, R. P. 2015. An improved workflow for quantitative N-terminal charge-based fractional diagonal chromatography (ChaFRADIC) to study proteolytic events in *Arabidopsis thaliana*. *Proteomics,* 15**,** 2458-69.

Vidmar, R., Vizovisek, M., Turk, D., Turk, B. & Fonovic, M. 2017. Protease cleavage site fingerprinting by label-free in-gel degradomics reveals pH-dependent specificity switch of legumain. *EMBO J,* 36**,** 2455-2465.

Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., . . . Ternent, T. 2015. 2016 update of the PRIDE database and its related tools. *Nucleic acids research,* 44**,** D447-D456.

von Stechow, L., Francavilla, C. & Olsen, J. V. 2015. Recent findings and technological advances in phosphoproteomics for cells and tissues. *Expert Rev Proteomics,* 12**,** 469-87.

Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J., Jr. 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl,* 44**,** 7342-72.

Wang, H., Chang-Wong, T., Tang, H. Y. & Speicher, D. W. 2010. Comparison of extensive protein fractionation and repetitive LC-MS/MS analyses on depth of analysis for complex proteomes. *J Proteome Res,* 9**,** 1032-40.

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

Weber, P., Harrison, F. W. & Hof, L. 1975. The histochemical application of dansylhydrazine as a fluorescent labeling reagent for sialic acid in cellular glycoconjugates. *Histochemistry,* 45**,** 271-277.

Wehr, N. B. & Levine, R. L. 2012. Quantitation of protein carbonylation by dot blot. *Anal Biochem,* 423**,** 241-5.

Weiss, A. & Stobo, J. D. 1984. Requirement for the coexpression of T3 and the T cell antigen receptor on a malignant human T cell line. *J Exp Med,* 160**,** 1284-99.

Weiss, A. & Stobo, J. D. 2015. Commentary: "The role of T3 surface molecules in the activation of human cells: a two-stimulus requirement for IL-2 production reflects events occurring at a pretranslational level". *Front Immunol,* 6**,** 163.

Weiss, A., Wiskocil, R. L. & Stobo, J. D. 1984. The role of T3 surface molecules in the activation of human T cells: a two-stimulus requirement for IL 2 production reflects events occurring at a pre-translational level. *J Immunol,* 133**,** 123-8.

Wejda, M., Impens, F., Takahashi, N., Van Damme, P., Gevaert, K. & Vandenabeele, P. 2012. Degradomics reveals that cleavage specificity profiles of caspase-2 and effector caspases are alike. *J Biol Chem,* 287**,** 33983-95.

Wells, J. M. & McLuckey, S. A. 2005. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol,* 402**,** 148-85.

Wiita, A. P., Seaman, J. E. & Wells, J. A. 2014. Global analysis of cellular proteolysis by selective enzymatic labeling of protein N-termini. *Methods Enzymol,* 544**,** 327-58.

Wildes, D. & Wells, J. A. 2010. Sampling the N-terminal proteome of human blood. *Proc Natl Acad Sci U S A,* 107**,** 4561-6.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., . . . Kuster, B. 2014. Mass-spectrometry-based draft of the human proteome. *Nature,* 509**,** 582-7.

Wilkins, M. R. & Appel, R. D. 2007. Ten years of the proteome. *Proteome Research.* Springer.

Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F. & Williams, K. L. 1996. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev,* 13**,** 19-50.

Williams, E. G., Wu, Y., Jha, P., Dubuis, S., Blattmann, P., Argmann, C. A., . . . Auwerx, J. 2016. Systems proteomics of liver mitochondria function. *Science,* 352**,** aad0189.

Wisniewski, J. R. 2017. Label-free and standard-free absolute quantitative proteomics using the "Total Protein" and "Proteomic Ruler" approaches. *Methods Enzymol,* 585**,** 49-60.

Wisniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. 2014. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics,* 13**,** 3497-506.

Wisniewski, J. R., Ostasiewicz, P., Dus, K., Zielinska, D. F., Gnad, F. & Mann, M. 2012. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol Syst Biol,* 8**,** 611.

Witus, L. S., Moore, T., Thuronyi, B. W., Esser-Kahn, A. P., Scheck, R. A., Iavarone, A. T. & Francis, M. B. 2010. Identification of highly reactive sequences for PLP-mediated bioconjugation using a combinatorial peptide library. *Journal of the American Chemical Society,* 132**,** 16812-16817.

Witus, L. S., Netirojjanakul, C., Palla, K. S., Muehl, E. M., Weng, C. H., Iavarone, A. T. & Francis, M. B. 2013. Site-specific protein transamination using N-methylpyridinium-4-carboxaldehyde. *Journal of the American Chemical Society,* 135**,** 17223-17229.

Wong, P. S. H. & Cooks, R. G. 1997. Ion trap mass spectrometry. *Current Separations,* 16**,** 85-92.

Wood, H. G. & Barden, R. E. 1977. Biotin enzymes. *Annu Rev Biochem,* 46**,** 385-413.

Wu, Y., Williams, E. G., Dubuis, S., Mottis, A., Jovaisaite, V., Houten, S. M., . . . Aebersold, R. 2014. Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell,* 158**,** 1415-1430.

Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. 2000. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom,* 35**,** 1399-406.

Xu, G. & Jaffrey, S. R. 2010. N-CLAP: global profiling of N-termini by chemoselective labeling of the alpha-amine of proteins. *Cold Spring Harb Protoc,* 2010**,** pdb prot5528.

Xu, G., Paige, J. S. & Jaffrey, S. R. 2010. Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat Biotechnol,* 28**,** 868-73.

Xu, G., Shin, S. B. & Jaffrey, S. R. 2009. Global profiling of protease cleavage sites by chemoselective labeling of protein N-termini. *Proc Natl Acad Sci U S A,* 106**,** 19310-5.

Yoshihara, H. A., Mahrus, S. & Wells, J. A. 2008. Tags for labeling protein N-termini with subtiligase for proteomics. *Bioorg Med Chem Lett,* 18**,** 6000-3.

Zeng, Y., Ramya, T. N., Dirksen, A., Dawson, P. E. & Paulson, J. C. 2009. High-efficiency labeling of sialylated glycoproteins on living cells. *Nat Methods,* 6**,** 207-9.

Zheng, J. & Bizzozero, O. A. 2010. Traditional reactive carbonyl scavengers do not prevent the carbonylation of brain proteins induced by acute glutathione depletion. *Free Radic Res,* 44**,** 258-66.

Zheng, S. & Doucette, A. A. 2016. Preventing N- and O-formylation of proteins when incubated in concentrated formic acid. *Proteomics,* 16**,** 1059-68.

Zhurov, K. O., Fornelli, L., Wodrich, M. D., Laskay, U. A. & Tsybin, Y. O. 2013. Principles of electron capture and transfer dissociation mass spectrometry applied to peptide and protein structure analysis. *Chem Soc Rev,* 42**,** 5014-30.

Zubarev, R. A., Zubarev, A. R. & Savitski, M. M. 2008. Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet? *J Am Soc Mass Spectrom,* 19**,** 753-61.