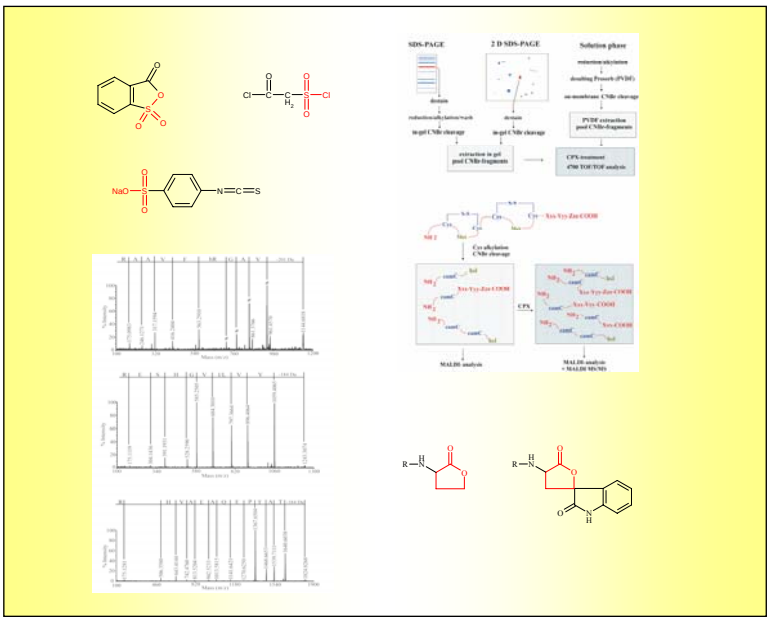
DEVELOPMENT OF NEW METHODS FOR C-TERMINAL AND DE NOVO SEQUENCE ANALYSIS WITH APPLICATION IN PROTEOMIC STUDIES



Kjell Sergeant

Thesis submitted to obtain the degree of Doctor (Ph.D.) in Sciences: Biochemistry

Promotor: Prof. Dr. J. Van Beeumen
Laboratory for Protein Biochemistry and Protein Engineering
University Ghent

2006

DEVELOPMENT OF NEW METHODS FOR C-TERMINAL AND DE NOVO SEQUENCE ANALYSIS WITH APPLICATION IN PROTEOMIC STUDIES

Kjell Sergeant

Thesis submitted to obtain the degree of Doctor (Ph.D.) in Sciences: Biochemistry

Promotor: Prof. Dr. J. Van Beeumen
Laboratory for Protein Biochemistry and Protein Engineering
University Ghent

2006

DANKWOORD

Op de eerste bladzijde van een doctoraat wordt doorgaans de appreciatie uitgesproken voor de steun die werd ondervonden tijdens het geleverde werk. Spontaan denk ik daarbij aan een citaat dat, zoals zovele citaten, verkeerdelijk, aan Newton wordt toegeschreven. De bekendste versie van dit citaat is de volgende: “If I’ve seen further, it is by standing on the shoulders of giants.”. Een oudere, langere en meer genuanceerde versie hiervan werd reeds omstreeks 1115 neergeschreven door een vroeg christelijk denker, John Of Salisbury: “We are like dwarfs sitting on the shoulders of giants. We see more, and things that are more distant, than they did, not because our sight is superior or because we are taller than they, but because they raise us up, and by their great stature add to ours.” Wanneer ik terugkijk op de laatste vijf jaar is het met de gevoelens die blijken uit dit citaat.

In de eerste plaats wil ik mijn promotor, Professor Van Beeumen, bedanken. U gaf me de kans op een moment dat er, met recht, kon getwijfeld worden of ik wel iets kon bereiken binnen het onderzoek. Ook na een eerste, fout gelopen, aanvraag voor een beurs heb ik geen aarzeling van uw kant ervaren. Van de reuzen, op wie hun schouders ik heb mogen staan, bent u waarschijnlijk de grootste en veel meer dan u bedanken voor de kansen en het vertrouwen kan ik daar niet tegenover stellen.

Iemand waar ik veel mee heb gediscussieerd is Bart Samyn, meestal waren dit vruchtbare gedachtenwisselingen. In ieder geval zou ik niet veel van wat hier beschreven staat hebben kunnen doen zonder de georganiseerde vrijheid waar jij op aandrong. Bij mij was die organisatie soms wat zoek, waarop jij mij meestal wel terug op spoor kreeg. Verder moet ik je zeker bedanken voor het nalezen en het in de plooi brengen van deze thesis. De meeste zinnen van langer dan 5 regels zijn eruit en alles is terug gebracht tot een taal die voor iedereen verstaanbaar is, merci.

Ik zou van deze gelegenheid ook willen gebruik maken om Professor Bart Devreese, die vanaf oktober de taak van professor Van Beeumen zal overnemen, te bedanken. U gaf mij de mogelijkheid om op korte tijd veel bij te leren in een studie van de klauwkikker, een studie die ik trouwens snel hoop af te sluiten met een artikel. Verder wil ik u alvast bedanken voor het vertrouwen dat u in mij stelt door op te treden als promotor voor toekomstige en huidige aanvragen.

Het “Instituut voor de Aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT-Vlaanderen), bedank ik voor de financiële steun, in de vorm van een specialisatiebeurs, die ik mocht ontvangen.

Ook wil ik de mensen van buiten het laboratorium die een belangrijke bijdrage hebben geleverd aan het werk dat hier staat beschreven bedanken. Professor Faro en Pedro Castanheira van de universiteit van Coimbra, voor het ter beschikking stellen van de stalen van cardosine A. Professor Swennen, Bart Panis en ‘de Seb’voor de samenwerking in verband met de identificatie van eiwitten geïsoleerd uit banaan. Ik hoop dat we deze samenwerking in de toekomst kunnen verder zetten. Ook de mensen van het ‘Laboratoire de protéomique’ (CPRGL, Luxemburg), Dr. Jenny Renaut en Dr. Hausman, wens ik via deze weg nogmaals te bedanken voor het uitvoeren van de automatisatie van onze methode.

De collega’s, het zijn er teveel om aan iedereen 5 regels te wijden. Tot mijn spijt zijn er een aantal die ik nog maar zelden zie. Elke, altijd hevige emoties maar ook altijd bereid tot

een gesprek en hulp bij niet altijd even geslaagde, ‘geheime’ experimenten. Frank, ik denk dat er niemand is met wie ik zoveel tijd heb doorgebracht met het bekijken van spectra, dankzij jou hebben fragmentatiespectra steeds minder geheimen. We hebben veel tijd gespendeerd met het bespreken van de toekomst en onze visie daarop. Ik hoop dat je in het werk dat je gaat doen in Zwitserland een deel van deze visie kan realiseren. Kris en Koen, we hebben niet veel samengewerkt, maar een gesprek over de wetenschap ging er altijd wel in. Griet, ik zou hier pagina’s kunnen voltypen met anecdotes en verhalen, maar ‘bedankt’ is al wat je eiste. Juffie, bedankt voor alles. Ondertussen zijn een aantal van deze mensen andere oorden gaan opzoeken, plaats latend voor een nieuwe garde massaspectrometristen die net als ik veel te bewijzen en dus weinig te verliezen hebben.

De resto-crew, ik heb genoten van de tijd aan tafel, al moet ik soms de indruk hebben gegeven dat ik niet snel genoeg weg kon zijn. Merci Bart, Jimmy, Paco en Samy. De andere collega’s, ik hoop dat ik niemand vergeet, Ann, Bjorn, Christiane, Debby, Dirk, Ester, Freddy, Gonzy, Ingrid, Isabel, Jan, John, Karen, Kenneth, Lina, Maarten, Savvas en Sofie wil ik bedanken voor de goede sfeer op het laboratorium en de constructieve uitwisseling van gedachten.

De 7 thesisstudenten die ik heb mogen begeleiden, dat ik niet altijd even vrolijk en mededogend was weet ik. Omdat jullie dat hebben verdragen en omdat jullie mij hebben geholpen in de uitwerking van dit werk: merci Griet, Bram de Rammelaere, Ruben, Stefanie, Pieter, Julie en Bram Miserez. Nu heeft enkel laatstgenoemde nog iets van mij te vrezen en hem wil ik extra bedanken omdat hij de laatste tijd vaak alleen verder heeft gemoeten.

Zonder de steun van mijn ouders, mijn broer, zus en natuurlijk de neefjes en nichtjes was ik misschien uit het oog verloren dat er belangrijker dingen zijn dan werk. Mijn ouders wil ik speciaal bedanken omdat ze me kansen hebben gegeven en me hebben laten falen zonder dat die een reden was om de steun te laten verzwakken. Tevens wil ik mijn schoonouders, mijn schoonzus en Wim bedanken voor de steun en hulp in de laatste acht jaar.

Als bijna laatste, maar belangrijkste, is er mijn vrouw. Katrien, ik weet niet of ik hier wel aan was begonnen zonder jou. Je hebt mij gesteund doorheen soms wilde ideeën en plannen. Ik hoop dat je trots bent op het geleverde werk al weet ik dat je niet 100% begrijpt waar het over gaat. Bovendien draag jij de laatste persoon die ik moet bedanken sinds enkele maanden nauw aan het hart. Als er iemand de laatste tijd heeft verhinderd dat ik begon te zweven of de belangrijke dingen vergat dan is zij het wel. Jouw geboorte werd aangekondigd voor midden juli, voor ons ben je echter al 7 maanden het belangrijkste feit. Hoe ik je hiervoor kan bedanken weet ik nog niet, maar daar hebben we het nog wel over wanneer je voor de eerste keer je zondagsgeld komt vragen.

Bedankt allemaal!

Kjell Bellem 2006

The best way to get a good idea is to
get a lot of ideas.

Pauling, Linus (1901-1994)

TABLE OF CONTENT

ABBREVIATIONS

AA	any amino acid
ACN	acetonitrile
ATH	alkylated thiohydantoin
CAF	chemically assisted fragmentation
CID/CAD	collision induced/activated dissociation
CNBr	cyanogen bromide
CPP	carboxypeptidase from the fungus Penicillum janthinellum
CPY	carboxypeptidase from Saccharomyces cerevisiae
CSAS	chlorosulfonylacetyl chloride
DE	delayed extraction
DIEA	N,N-diisopropylethylamine
DTT	dithiothreitol
2D-PAGE	two-dimensional polyacrylamide gel electrophoresis
ESI	electrospray ionization
EST	expressed sequence tag
FAB	fast atom bombardment
FTMS	Fourier transform mass spectrometry
hR	homo-arginine (homo-Arg)
hsl	homoserine lactone
IAA	iodoacetamide
ICAT	isotopically coded affinity tag
IEF	isoelectric focusing
IPG	immobilized pH gradient
LC	liquid chromatography
LysC	lysyl endoprotease
m/z	mass-to-charge ratio
MALDI	matrix-assisted laser desorption/ionization
MDLC	multi dimensional liquid chromatography
MQ	Milli-Q water
MS	mass spectrometry
MS/MS	tandem mass spectrometry
Mw	molecular weigth
PAGE	polyacrylamide gel-electrophoresis
PCR	polymerase chain reaction
PIC-model	pathways in competition model
PMF	peptide mass fingerprinting
PMOC	proopiomelanocortin
pmol	picomol
PSD	post source decay
PTM	posttranslational modification
Q-TOF	quadrupole time of flight
RP	reverse phase

SACA	2-sulfobenzoic acid cyclic anhydride
SCX	strong cation exchange
SDS	sodium dodecyl sulfate
SPITC	4-sulfophenyl isothiocyanate
TFA	trifluoroacetic acid
THF	tetrahydrofuran
TIS	timed-ion-selector
TMPP	tris(trimethoxy)phosphonium acetate
TOF	time of flight

Symbols and residual masses of the 20 common amino acids

Amino acid	Symbol		Monoisotopic (Da)	Average (Da)
Alanine	Ala	A	71.0371	71.0788
Arginine	Arg	R	156.1011	156.1876
Asparagine	Asn	N	114.0429	114.1039
Aspartic acid	Asp	D	115.0269	115.0886
Cysteine	Cys	C	103.0092	103.1448
Glutamine	Gln	Q	128.0586	128.1308
Glutamic acid	Glu	E	129.0426	129.1155
Glycine	Gly	G	57.0215	57.0520
Histidine	His	H	137.0589	137.1412
Isoleucine	Ile	I	113.0841	113.1595
Leucine	Leu	L	113.0841	113.1595
Lysine	Lys	K	128.0950	128.1742
Methionine	Met	M	131.0405	131.1986
Phenylalanine	Phe	F	147.0684	147.1766
Proline	Pro	P	97.0528	97.1167
Serine	Ser	S	87.0320	87.0782
Threonine	Thr	T	101.0477	101.1051
Tryptophan	Trp	W	186.0793	186.2133
Tyrosine	Tyr	Y	163.0633	163.1760
Valine	Val	V	99.0684	99.1326

LIST OF PUBLICATIONS

I De Clerck E, Gevers D, Sergeant K, Rodriguez-Diaz M, Herman L, Logan NA, Van Beeumen J, De Vos P.
Genomic and phenotypic comparison of Bacillus fumarioli isolates from geothermal Antarctic soil and gelatine. *Res Microbiol* (**2004**) 155(*6*):483-90.

II Samyn B, Debyser G, Sergeant K, Devreese B, Van Beeumen J.
A case study of de novo sequence analysis of N-sulfonated peptides by MALDI TOF/TOF mass spectrometry. *J Am Soc Mass Spectrom* (**2004**) 15(*12*):1838-52.

III Samyn B, Sergeant K, Castanheira P, Faro C, Van Beeumen J.
A new method for C-terminal sequence analysis in the proteomic era. *Nat Methods* (**2005**) 2(*3*):193-200.

IV Castanheira P, Samyn B, Sergeant K, Clemente JC, Dunn BM, Pires E, Van Beeumen J, Faro C.
Activation, proteolytic processing, and peptide specificity of recombinant cardosin A. *J Biol Chem* (**2005**) 280(*13*):13047-54.

V Sergeant K, Samyn B, Debyser G, Van Beeumen J.
De novo sequence analysis of N-terminal sulfonated peptides after in-gel guanidination. *Proteomics* (**2005**) 5(*9*):2369-80.

VI Samyn B, Sergeant K, Memmi S, Debyser G, Devreese B, Van Beeumen J.
MALDI TOF/TOF de novo sequence analysis of 2D-PAGE separated proteins from *Halorhodospira halophila*, a bacterium with unsequenced genome. *Electrophoresis* (**2006**), in press.

VII Samyn B[*], Sergeant K[*], Carpentier S, Debyser G, Panis B, Swennen R, Van Beeumen J.
Homology-based functional proteome analysis: a successful approach for the non-model plant *Musa* spp. *Mol Cell Proteomics* (**2006**), submitted.
[*] These authors contributed equally

I am among those who think that science has great **beauty**. A scientist in his laboratory is not only a technician: he is also a **child** placed before natural phenomena which impress him like a **fairy tale**.

Curie, Marie (1867-1934)

Foreword

On the 28th of February 1953, Francis Crick, while entering a pub in Cambridge, proclaimed "We have found the secret of life". Although this is probably too far fetched, with the elucidation of the structure of DNA, Watson and Crick realized one of the most important breakthroughs in 20th century science. In those days the DNA-molecule was considered as a molecule that was hard to handle. With the developments of techniques of what is known as molecular biology or biotechnology, it is now probably the most studied and best understood biopolymer. It can be isolated, multiplied, specific parts can be tagged at very high sensitivity and interesting sequences can be specifically transferred to other organisms and their products studied. Most importantly, the sequence of nucleotidemonomers can be determined fully automated at rates of thousands a day. The first complete viral genomes were determined in the late 1970s (Fiers *et al*, 1978), and a first prokaryotic genome in 1995 (Fleischmann *et al*, 1995). Thirteen years after the inauguration of the Human Genome Project, these efforts culminated in the submission of a completed human genome in public sequence databases (Carroll, 2003). These steps were important milestones in the recent development of biological science and show the potential of large-scale shotgun genome analysis. Today, more then 300 genomes are sequenced and, for some 1500 more, sequencing is in progress (www.genomesonline.org). Although the development of improved technology continues, striving towards 'personal genomics' (Chan, 2005), the central debate has shifted from 'how?' to 'whose next?' (Gewolb, 2001a; Gewolb, 2001b). In the 1990s genomics has changed from an essentially academic exercise to a serious commercial endeavour, thereby marking the advent of a new scientific era, post-genomics.

The impact of the availability of genomic sequences in public databases extends far beyond mere information exchange. Prompted by the accessibility of genome sequences, the paradigm of life sciences shifted from small-scale reductionistic studies to large-scale holistic analysis approaches. It was realized that the genome of an organism only represents a first layer of information; in order to describe the properties of biological systems the study of gene/protein expression was required. Although analysis of transcription provides a first clue on the dynamic properties of a cell or tissue, the dynamism that is inherent to life can only be adequately assessed at the protein level. Proteins are the key functional molecules in living systems and are involved in all cellular processes as catalysts, regulators or structural scaffolds. To perform all these functions the expression and activity of proteins is strictly regulated and dependent on temporal and environmental influences. Therefore, the complete characterization, both qualitative and quantitative, of all proteins present in a cell or tissue at a certain time and under specific environmental conditions is currently the focal point of scientific efforts aimed at the comprehension of cellular life.

The understanding of all mechanisms that govern life at all levels is explicitly the aspiration of 'Systems Biology' (Medina, 2005). A systems level characterization addresses three main questions. First, what are the parts of the system? Second, how do the parts work? Third, how do the parts work together to accomplish a task? Systems Biology is driven by computational analysis of large data sets (Aderem, 2005). Iterative cycles of data collection, modeling, hypothesis formulation and testing finally result in model refinement and further cycles. To accomplish the goal of Systems Biology, complete data sets describing the identity, the quantity and the properties of all the molecules that compose an organism are mandatory. Such data sets can be generated for DNA and mRNA; automated DNA sequencers enable the sequencing of genomes and microarray analysis permits global transcription profiling. However, the sheer chemical and physical complexity and the dynamism that is observed at

the protein and the metabolite level currently preclude a similar in-depth analysis of these molecules.

In the work presented here, we focus on the development of new techniques to solve some of the remaining problems in the study of the proteome. After a general outline of the large domain of biological sciences with the emphasis on the 'omics'-sciences, an overview is given of the most frequently used analytical methods in proteomics (Part 1). Proteomics is a highly dynamic and competitive field; therefore a complete literature overview is beyond the scope of this work. Rather an impression of the most relevant topics is given. References to each part are given at the end of the respective chapters.

Despite the availability of numerous techniques to study proteins, hiatuses remain. In Part 2, we show results obtained by using our new technique for C-terminal sequence analysis of proteins. After a general overview of the related literature (Part 3.1), the implementation of in-gel guanidination in a protocol for N-terminal sulfonation of peptides, a technique aimed at facilitating *de novo* sequence determinations from fragmentation spectra, is described in Part 3.2. The results of the use of this improved protocol for the identification of proteins from organisms for which no genomic information is available, together with initial results of the automation of this protocol, are described Part in 3.3. Conclusions and future developments are gathered in Part 4.

References

Aderem, A. (**2005**). Systems biology: its practice and challenges. *Cell* **121**(4): 511-3.

Carroll, S.B. (**2003**). Genetics and the making of Homo sapiens. *Nature* **422**(6934): 849-57.

Chan, E.Y. (**2005**). Advances in sequencing technology. *Mutat Res* **573**(1-2): 13-40.

Fiers, W.; Contreras, R.; Haegemann, G.*; et al.* (**1978**). Complete nucleotide sequence of SV40 DNA. *Nature* **273**(5658): 113-20.

Fleischmann, R.D.; Adams, M.D.; White, O.*; et al.* (**1995**). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**(5223): 496-512.

Gewolb, J. (**2001a**). Genomics. Animals line up to be sequenced. *Science* **293**(5529): 409-10.

Gewolb, J. (**2001b**). Genome research. DNA sequencers to go bananas? *Science* **293**(5530): 585-6.

Medina, M. (**2005**). Genomes, phylogeny, and evolutionary systems biology. *Proc Natl Acad Sci U S A* **102 Suppl 1**: 6630-5.

PART 1

GENERAL INTRODUCTION

1.1. Life sciences

The ultimate goal of biological sciences is the description of the principles and the rules that control biological processes. Starting from biological samples, this can be done at different levels. The study of the phylogenetic interrelations between species, their phenotypes and their morphology is the topic of pure biology. The focus of biology therefore is on the organism as a whole and on its interactions with the environment. While biology is mainly descriptive, physiology and 'Systems Biology' are broader: studying organisms as aggregates of molecules that interact to produce a response to biological phenomena or to allow an organism to sustain in a dynamic environment. Although physiology and 'Systems Biology' are sometimes considered to be synonymous, the difference is situated in the strong computer-based approach of 'Systems Biology'. Together with ecology, the study of the different biological and anorganic entities that are the components of a environment and their interactions, these three scientific areas of expertise provide the questions that are answered in more specialized scientific fields.

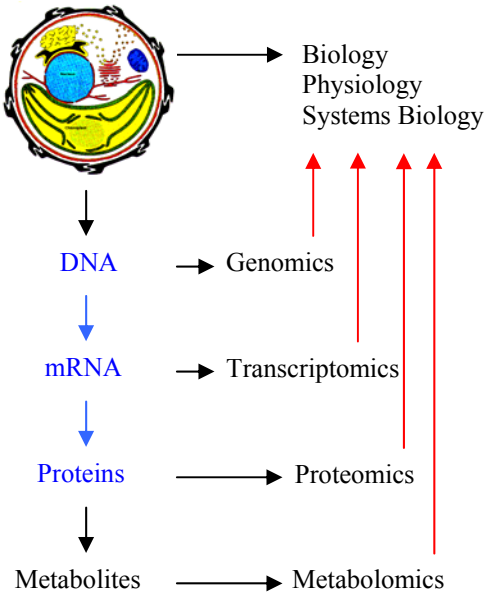


Figure 1.1. Overview of the different scientific disciplines discussed in this section. Indicated in bleu is the central dogma of life, the flow of biological information in a cell. Red arrows indicate the flow of experimental data, finally resulting in a synthesis at the organisomal level.

Following the central dogma of life (Figure 1.1), the flow of biological information from DNA, via mRNA to proteins, specific scientific disciplines have emerged. Nearly all information that is transferred to offspring is encoded in an organism's DNA. The DNA blueprint of an organism is inherently stable over short time-spans and thus little flexibility to respond to the changing environment is provided. To be able to respond to changes, information stored in the DNA is transcribed in mRNA. In prokaryotes, flexibility is added in this process through a strict qualitative and quantitative control of transcription. On the contrary, in eukaryotic organisms processing of the primary transcripts adds a further level of control and flexibility. Although RNA is an effector molecule in some processes, e.g. tRNA and rRNA, the number of functions that can be performed by the relatively simple molecules is limited. Only after translation of the nucleotide sequences in amino acid sequences, the flexibility that is required to reside in a dynamic environment is possible.

1.1.1. Genomics

Although high-throughput sequencing of genomes is by far the most appealing feature in genome research, genomics encompasses a broad range of disciplines that study the information encoded in the DNA of an organism. Meldrum reviewed the approaches to automated genomics and described the equipment used for automated genome sequencing (Meldrum, 2000a; Meldrum, 2000b). Despite the growing number of sequenced genomes, some clades are unrepresented among the fully sequenced organisms. Because of their

relatively small genome, bacterial organisms make up more than 80% of the 340 completed genomes (www.genomesonline.org update 29/01/06), while Viridiplantae are most prominently underrepresented.

The focus in genomic research has now shifted from the acquisition of data to the extraction of useful information from nucleotide sequences. A first step in this process is the prediction and annotation of genes. Different methods for gene annotation are commonly used (Stein, 2001) and new methods of growing complexity are regularly described. Two different approaches for gene annotation exist (Borodovsky *et al*, 1994), an extrinsic homology-based one and an intrinsic one that uses computational analysis to predict gene sequences based on known properties of coding-sequences (Pavy *et al*, 1999; Rouze *et al*, 1999). Homology-based gene annotation is the fastest and most reliable method. Nevertheless, only half of the genes can be annotated using this approach, predictive models are required to annotate the remainder. Apart from the nucleotide sequence alone, the increased knowledge of the physical arrangements of genes in chromosomes and of the changes of genes over an evolutionary time-span allows the addition of multiple dimensions for genome annotation. Further studies are headed to incorporate this information in programs for gene annotation (Taher *et al*, 2004; Reed *et al*, 2006). Nevertheless, a large percentage (40%) of the human genes remains without functional annotation (Bateman *et al*, 2002). The problems associated with genome annotation are illustrated in some articles published in 2003 (Oliver *et al*, 2003). Using a very loose gene finding algorithm about 22000 possible genes were found in *Drosophila*, 50% more than in the best annotated *Drosophila* database at that time (release 3.1; http://flybase.bio.indiana.edu/). At least half of the predicted genes that were only found using the loose prediction protocol were confirmed as expressed RNAs (Hild *et al*, 2003).

A genome is a repository of information that only reflects the potential of an organism. It thus only presents a first layer of complexity. Therefore, it was clear that new techniques were required to study the functional complement of genomes. Genome transcripts, gene products and metabolites are the topic of 'functional genomic' studies. As the study of the genome is known as 'genomics', the study of the different functional complements of the genome are now recognized as '-omics' sciences.

1.1.2. Transcriptomics

While the genome of an organism is inherently stable, temporal flexibility is added at the transcriptome level. In prokaryotes, the transcriptome is mainly regulated by controlled transcription and degradation of transcripts. In eukaryotes, on the contrary, a single gene can result in different transcripts through processes such as alternative splicing, the use of alternative promoters and RNA-editing.

This environment-depending variation allows the study of cell-specific gene expression and thus a first molecular clue on the functional properties of the cell/tissue. Today, profiling of the transcriptome is done using high-throughput micro-analysis systems. To determine which genes are activated in various circumstances, a comparative differential approach is used (Cekan, 2004). In this approach, expression profiles from different states are compared, allowing to isolate the genes expressed under specified conditions. The use of transcript-profiling to study breast cancer was described. A transcript-profile from each of 65 tumors was obtained using cDNA microarrays representing > 8000 human genes (Perou *et al*, 2000). This information was subsequently used to predict the outcome of different treatments on every tumor. Furthermore, clustering of the activated genes allows establishing groups of

genes that are generally activated simultaneously. The clustering of co-expressed genes, via the 'guilt by association'-principle (Eisen *et al*, 1998; Oliver, 2000; Wolfe *et al*, 2005), and the study of their expression in different tumors allowed to distinguish tumor subclasses (Sorlie *et al*, 2001).

Although gene expression profiling allows the study of the genes that are expressed under specific physiological cell states, no information on the effective concentration of active gene products is available from these studies, nor can an explanation for the absence/increase of expression be deduced from these profiles.

1.1.3. Proteomics

Proteins are the main effector molecules encoded by the genome of a cell and are involved in every aspect of the cellular function as structural scaffolds, catalysts in both simple and complex cellular processes, and signal transducers (Pandey *et al*, 2000). However, the original definition of the term proteome '**Prote**in complement encoded by a gen**ome'** does not reflect the dynamic properties of what is now known as the proteome (Wilkins *et al*, 1996). A more comprehensive definition could be 'proteomics is the study of all proteins expressed at a given point in time under given circumstances by a specific type of cell or tissue'.

Variation at the protein level is conveyed by different processes; including co- and posttranslational modifications of amino acid side chains, proteolytic processing, transport of proteins across membranes, protein interactions and the action of inteins. Therefore, one gene can encode for numerous proteins, resulting in the economical use of coding sequences and the specific activation/deactivation of pathways.

Proteomics is hampered by the sheer complexity of the proteomes of cells and tissues. Furthermore, in contrast to nucleotides, which can be amplified by the well-known 'polymerase chain reaction' (PCR) technique, proteins can not be amplified. Therefore, the protein composition of samples reflects the cellular protein composition, in which the dynamic range of individual proteins often spans several degrees of magnitude. Currently, no techniques are available that allow the complete profiling of the proteome. This lack of comprehensive methods is mainly due to the highly diverse nature of proteins, from soluble (hydrophilic) to membrane-inserted (hydrophobic) and from small (< 50 amino acids) to very large (> 4000 amino acids). Currently, no methods are available that allow quantitative proteome profiling with a throughput comparable to that obtained at the transcriptome level. Furthermore, the currently used high-throughput techniques for protein identification are based on meticulous matching of peptide or fragment masses to mass spectra generated by in-silico processing of database entries, limiting their use to organisms for which sequence information is available. Most prominently, the characterization of posttranslational modifications and the identification of low abundant proteins is difficult with current techniques. Recently, a number of reviews were published in which the remaining pitfalls in proteome research are discussed (Reinders *et al*, 2004; Rose *et al*, 2004; Alaiya *et al*, 2005; Bertone *et al*, 2005; Garbis *et al*, 2005).

1.1.4. Other '–omics'-sciences

The metabolome is made up of all of the low-molecular weight molecules (metabolites) present in a cell at a particular time, and their levels can be regarded as the

functional response of biological systems to genetic or environmental stimuli (Fiehn, 2002; Rochfort, 2005). The analysis of the metabolome is particularly challenging due to the diverse chemical nature of metabolites. As for proteins, the highest diversity in metabolites is found in eukaryotic organisms with hundreds of thousands of metabolites estimated to occur in plants (Fridman *et al*, 2005). Although the real strength of metabolomics can only be realized in combination with the identification of the proteins and genes responsible for the synthesis of a metabolite, no genomic information is a priory required to profile the metabolome of species. At present, as for proteomics, a standard method is not established for metabolomics (Fukusaki *et al*, 2005).

A large number of new '-omics' terms were recently proposed to coin specialized subdisciplines. Examples of proteomic subdisciplines include phosphoproteomics (Mukherji, 2005), secretomics (Tjalsma *et al*, 2000), peptidomics (Verhaert *et al*, 2001; Baggerman *et al*, 2004) and interactomics (Cusick *et al*, 2005), in which respectively phosphorylated proteins, secreted proteins, functional peptides and the interaction between proteins are studied. Because proteins generally function in aggregates, interactomics is probably the discipline most relevant to functional genomics. In the study of protein interactions the 'guilt by association principle', used in the clustering of transcripts, is complemented by the 'majority rule'. This rule states that when the majority of proteins in a complex are involved in the same global function, the other, unknown partners are also involved in the same function. Methods used to unravel protein-networks include yeast-2-hybrid screening (Walhout *et al*, 2001) and affinity-purification combined with mass spectrometry (Bauer *et al*, 2003).

1.2. Proteomics

Nearly 100 years after the first purification of proteins in the late 19th century, large-scale protein analysis, proteomics, became possible. Nonetheless, the first complete proteome is not composed yet. Above all, these advances are technology-driven, and new scientific frontiers were always set with the implementation of improved analysis techniques. For instance, 10 years ago, one protein was identified for each 2D-gel spot (Shevchenko *et al*, 1996). More recently, the reliable identification of multiple proteins in one gel spot has become the standard.

The answer to the question "what is hampering the scientific community to study proteomics to its full potential?" is multiple. But all arguments stem from the inherently diverse and complex properties of proteins themselves. Firstly, DNA and RNA are hydrophilic molecules that are only stable in a few conformations and found to be solubilized in the cytoplasm in prokaryotes and in the nucleus in eukaryotes. Proteins, on the other hand, are distributed over all cell structures. The study of water-soluble cytoplasmic proteins requires approaches that are different from the extraction procedures of membrane-associated or integral membrane proteins. Secondly, the study of proteins is severely hampered because of the lack of an amplification method comparable to the polymerase chain reaction (PCR) for poly-nucleotides. Because of the lack of an amplification method, the protein composition in the analyzed sample reflects the relative protein abundances in the organism. Thirdly, cellular protein extracts contain a large number of different proteins, a sample complexity that, mainly in eukaryotes, is much higher than the complexity seen at the transcriptome level. Therefore, efforts to characterize the total protein complement of cells require extensive separation prior to analysis. Using bottom-up approaches, the sample complexity increases further by a factor of about 50. Finally, the study of proteomics is limited by the lack of knowledge of the fragmentation behaviour of proteins and peptides. This precludes the development of accurate

predictive models and results in the use of algorithms for database comparison that only use a part of the information available in mass spectra.

These limitations are apparent from the difficulties associated with the use of microarrays for protein identification, an approach that is now routinely used for DNA and cDNA research. Protein microarrays are miniaturized arrays of antibodies, on which fluorescently labeled sample is applied (Haab *et al*, 2001) or that are screened using sandwich assays. However, using these approaches, the detected signal is not a simple function of protein abundance (Barry *et al*, 2003). Because the strength of protein-antibody interactions depends on the 3D-structure of proteins and on the presence/absence of posttranslational modifications, the affinity range of these interactions is broad and unpredictable. Furthermore, the use of antibody arrays excludes the study of membrane proteins since these can not be solubilized without disrupting their structure. As a result, absolute protein quantification with microarrays is not yet possible and, in practice, the use of antibody microarrays is limited to study dedicated subjects in comparative and competitive approaches (Barry *et al*, 2004).

1.2.1. Approaches

1.2.1.1. Overview

A summary, representing the most frequently used techniques is given in Figure 1.2. New methods are generally the result of developments whereby one of the depicted steps is varied. In essence, the outline remained unchanged since the technique of 2D gel electrophoresis combined with endoprotease digestion and mass spectrometric analysis of the resulting peptides was described for the first time in 1993 (Part 1.2.2.1). After protein extraction, proteins or their corresponding peptides are separated and the resulting fractions are analyzed by mass spectrometry. Critical steps in this scheme are the extraction and the separation. More recently, methods to simplify the sample mixtures were introduced at the protein or at the peptide level, as indicated in the scheme (respectively Fractionation 1& 2).

The extraction of a representative protein sample is a first bottleneck in the study of proteomes. While some groups of proteins, notably those composing the secretome, are fairly easy to gather, proteins from most cells or tissues require the use of extraction solvents containing high concentrations of detergents and/or chaotropes. The selective enrichment of proteins using one extraction method compared to the next (Harder *et al*, 1999; Carpentier *et al*, 2005) precludes the foundation of a standard method. Therefore, stepwise extraction protocols have been described (Bunai *et al*, 2005). Recently, several methods were published wherein protein extraction is avoided, for instance by performing digestions on membrane proteins still inserted in the membrane (Kuhn *et al*, 2003; Nielsen *et al*, 2005).

After protein extraction, different approaches are available to lower the complexity of a sample (Fractionation 1 in Figure 1.2). By performing a fractionation at the protein level, the peak capacity of the overall method increases. This step is often referred to as prefractionation. Techniques such as chromatofocusing (ion exchange chromatography using buffers of constant molarity but changing pH) (Kang *et al*, 2000; Kang *et al*, 2004) and Serial Isoelectric Focusing (SIEF) (Farhoud *et al*, 2005) are used to subdivide the sample in a number of fractions. A review published in 2005 discusses several of the different approaches used for isoelectric prefractionations, including the Rotofor, a commercially available device from BioRad (Righetti *et al*, 2005b). Other approaches for prefractionation include the use of immunoaffinity to deplete biological samples from highly abundant proteins, e.g. albumin and

Ig's from blood samples (Duan *et al*, 2005). Although kits for depletion of albumin and Ig's are commercially available, the utility of this approach was recently questioned because aspecific depletion of other proteins was observed (Granger *et al*, 2005). Selection of specific groups of proteins can be done using lectins for the selection of glycoproteins (Madera *et al*, 2005), immobilized ligands for the selection of binding proteins, or the use of immobilized proteins for the selection of interacting proteins. Most of these approaches have been reviewed in 2005 by Righetti *et al* (Righetti *et al*, 2005a). Finally, the establishment of procedures for the isolation of subcellular proteomes provides an approach to dissect the total proteome (Jung *et al*, 2000).

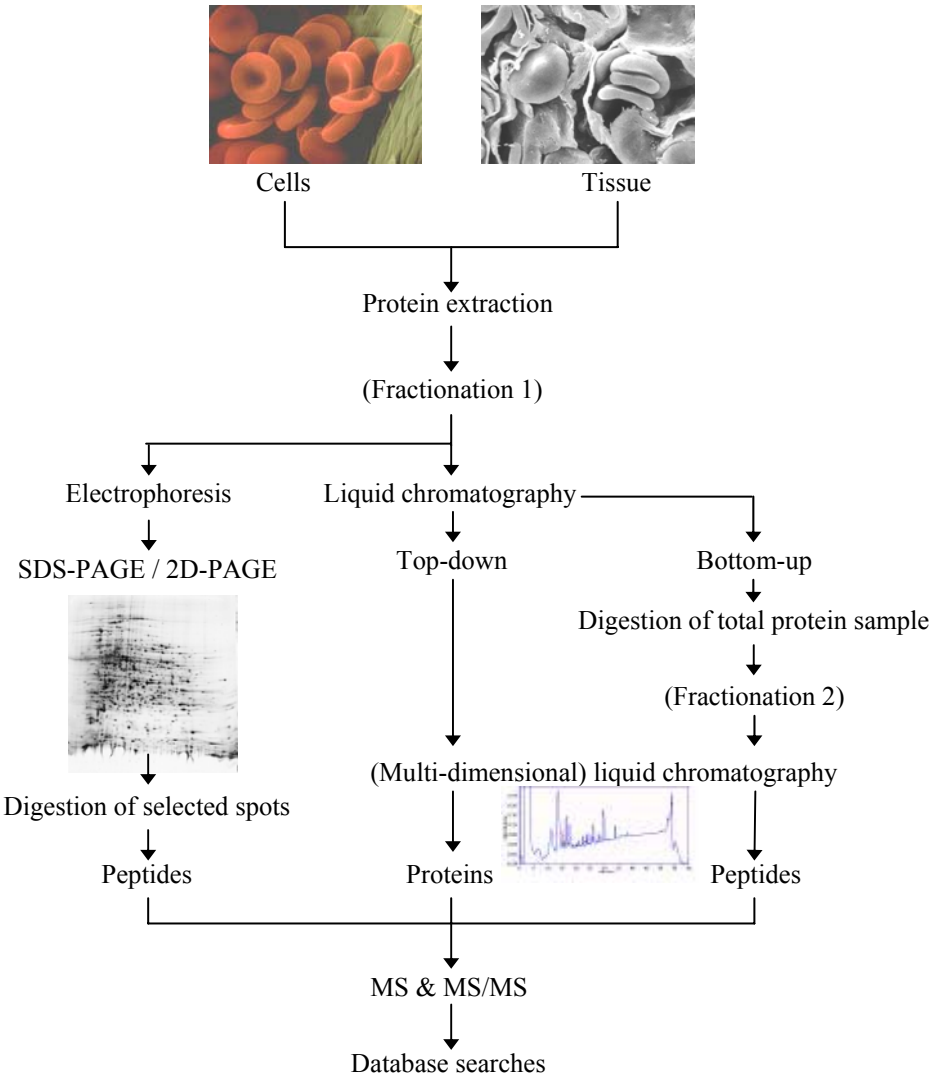


Figure 1.2. General overview of the different currently used approaches. Samples are extracted from cells or tissues; starting from this protein sample, either a top-down approach, i.e. the separation of intact proteins, or a bottom-up approach, i.e. the separation of peptides after digestion of the protein sample, is performed. Each of the steps can be varied and different fractionation steps can be inserted. The different steps are briefly discussed in the text.

Separation of proteins using 2D-gel electrophoresis is the oldest technique in proteomics. In the 1970's, before the term proteomics was invented, 2D gel electrophoresis was already used to study the cellular proteome and to build protein databases (O'Farrell, 1975). In 2D-PAGE, proteins are separated by isoelectric focusing in the first dimension and by PAGE in the second dimension. Because these separations are based on orthogonal parameters, respectively isoelectric point and molecular weight, 2D-PAGE results in a high

peak capacity (Gorg *et al*, 2000; Rabilloud, 2002; Gorg *et al*, 2004). Variations on the combination of isoelectric focusing in the first dimension with SDS-PAGE as the second dimension are described (Schagger *et al*, 1991; Oh-Ishi *et al*, 2000). The main disadvantages of 2D-gels are the limited reproducibility, failure to resolve very large and very small proteins, and the inability to separate most membrane proteins. Furthermore, 2D-PAGE performs equally poor in the separation of extremely acidic and basic proteins, as well as in the visualization of low abundant proteins (Peng *et al*, 2001). However, approaches to overcome these limitations have been proposed (Santoni *et al*, 2000; Luche *et al*, 2003; Bunai *et al*, 2005). Because a 2D-gel provides a strong, attractive image of intact proteins, which reflects changes in protein expression level, isoforms or post-translational modifications, 2D-PAGE remains the workhorse for proteomics (Hecker, 2005).

In the bottom-up approach, complete protein extracts, or fractions thereof, are digested, generally using trypsin, and the separation of proteins is replaced by the separation of the corresponding peptides. Tryptic peptides are more homogeneous than proteins, displaying a limited mass distribution and generally containing two basic groups, the N-terminal amino group and the C-terminal ε-aminogroup of lysine or the guanidino group of arginine. Although the sample complexity is increased, the homogeneity allows automated, standardized separations using (multi-dimensional) liquid chromatography (MDLC). Multiple methods have been described to select specific peptides out of these mixtures (Fractionation 2 in Figure 1.2). These approaches strive towards the dogma “one protein, one peptide” and are often coupled to the relative quantification of the proteins (Zhang *et al*, 2004). One of the first methods devised in this respect is Isotope-Coded Affinity Tag (ICAT) (Gygi *et al*, 1999a). Modification of cysteine with a biotin-containing reagent allowed isolation of cysteine containing peptides using an avidin-biotin affinity purification strategy. However, not every protein contains cysteine residues and most proteins have more than one. Only selection of either the N- or the C-terminal peptide of proteins allows the most rigorous simplification of samples. The use of anhydrotrypsin to select C-terminal peptides is discussed in the section on C-terminal sequence analysis (Part 2.1). The specificity accomplished in N-terminal derivatization makes the N-terminal peptide a more likely candidate for selection, and studies to attain such approach have been performed (McDonald *et al*, 2005). Gevaert *et al* described a platform that allows the specific selection of peptides. Cofradic, **Co**mbined **Fra**ctional **Di**agonal **C**hromatography, uses chemical modifications to attain shifts in retention time during consecutive reversed phase (RP) separations. This method has been applied for the specific isolation of N-terminal and methionine-containing peptides (Gevaert *et al*, 2003; Gevaert *et al*, 2005).

As an alternative to 2D-PAGE, methods have been developed that use MDLC to separate peptides mixtures after digestion of cellular protein extracts. After multidimensional separation, the eluate can be either analyzed online by ESI-MS, or alternatively, the fractions are spotted on MALDI targets plates in LC-MALDI-MS approaches (Hattan *et al*, 2005). Although the combination of various chromatographic techniques has already been applied successfully, the most frequently used combination is the hyphenation of strong cation or anion exchange (SCX/ACX) with RP chromatography (Issaq *et al*, 2005). In MudPIT-analysis, **Mu**lti**d**imensional **P**rotein **I**dentification **T**echnology, biphasic columns are packed in the capillary ion source of ESI-MS. Peptide mixtures are loaded and stepwise eluted from the SCX-phase onto the RP-packing. A gradient, with an increasing concentration of organic solvent, is formed and peptides are eluted directly into the mass spectrometer (Washburn *et al*, 2001). In other applications the two dimensions are separated, either online using trapping columns (Nagele *et al*, 2003), or offline (Vollmer *et al*, 2003). Different combinations of

chromatographic phases were reviewed in 2003 (Wang *et al*, 2003). The orthogonality of different combinations of chromatographic separations, as a measure for their maximal peak capacity, was recently studied (Gilar *et al*, 2005). Despite all these developments, a study on myelin sheets recently demonstrated that, in order to identify a maximum number of proteins, combinations of MDLC and 2D-PAGE are required. The use of each technique resulted in the identification of proteins not identified by the other method (Vanrobaeys *et al*, 2005).

1.2.1.2. Mass spectrometry

The main (r)evolution that enabled proteomics was the development of new mass spectrometric techniques. Mass spectrometry was pioneered in the beginning of the 20th century by Thompson and Aston for the analysis of the isotopic composition of elements using magnetic sector instruments. New mass analyzers such as 'time-of-flight'-tubes (TOF), quadrupoles and ion traps were developed in the 1950's by, among others, Wolfgang Paul. At that time, analysis of proteins and peptides (or biomolecules in general) was troublesome since mass spectrometry requires the formation of charged, gaseous molecules. Some 30 years later the introduction of new, soft, ionization techniques allowed the use of MS for the analysis of proteins and peptides at high sensitivity. The development of electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI) mass spectrometry by respectively John Fenn and Koichi Tanaka would soon prove to be of pivotal importance for proteome research.

Basically, a mass spectrometer is composed of three compartments; an ion source, a mass analyzer and a detector device. The mass analyzer can be composed of highly variable parts and most modern mass spectrometers have more than one mass analyzer. During this work only one type of mass spectrometer was used, the Applied Biosystems 4700 TOF/TOF. Therefore, only MALDI ionization and TOF analysis will be described, whereas the principle of ESI will be discussed briefly.

Instrumentation

Electrospray ionization

During electrospray ionization, the sample is introduced in a solvent and ions are generated at atmospheric pressure. Although in the 1960's large molecules were already introduced in the gas-phase, the application of electrospray for biological mass spectrometry was only founded in 1989 by the group of John Fenn (Fenn *et al*, 1989). Initially, solvent containing analytes was forced through a conducting capillary at high potential (Smith *et al*, 1990). For positive mode analysis the solvents are acidic and the potential on the capillary positive, thus protons are repulsed from the tip of the capillary, forming a 'Taylor cone'. Charged droplets are ejected from the Taylor cone once repulsion between equal charges becomes higher than the surface tension of the solvent. These charged droplets move towards the counter electrode situated in front of the mass analyzer (Kebarle *et al*, 1997). While the nebula approaches the mass spectrometer, solvent evaporates from the droplets, facilitated by a stream of drying gas. The loss of volume due to evaporation results in an increased charge density on the surface of the droplets, and when the charge repulsion exceeds the surface tension of the droplet, the so called 'Rayleigh limit', the droplets break up (Figure 1.3) (Kebarle *et al*, 1993). Because analytes are directly ionized from liquids, ESI is currently mainly used in LC-MS settings

While there is consensus on the ionization process up to this point, the final steps in the ionization process are still under study (Fenn, 1993; Kebarle, 2000). Two models that describe the final step in the ionization during ESI, the 'charged residue model' (Dole *et al*, 1968) and the 'ion evaporation model' (Iribarne *et al*, 1976), have been proposed. In 1996, a new type of ESI source was presented, the nanoelectrospray (Wilm *et al*, 1996). The ionization in nanoelectrospray is purely based on electrostatics. During nanoelectrospray, smaller droplets are formed than during conventional ESI, 180 nm compared to 1µm (Juraschek *et al*, 1999), resulting in a higher sensitivity and a higher tolerance for salts in the sample (Griffiths, 2000).

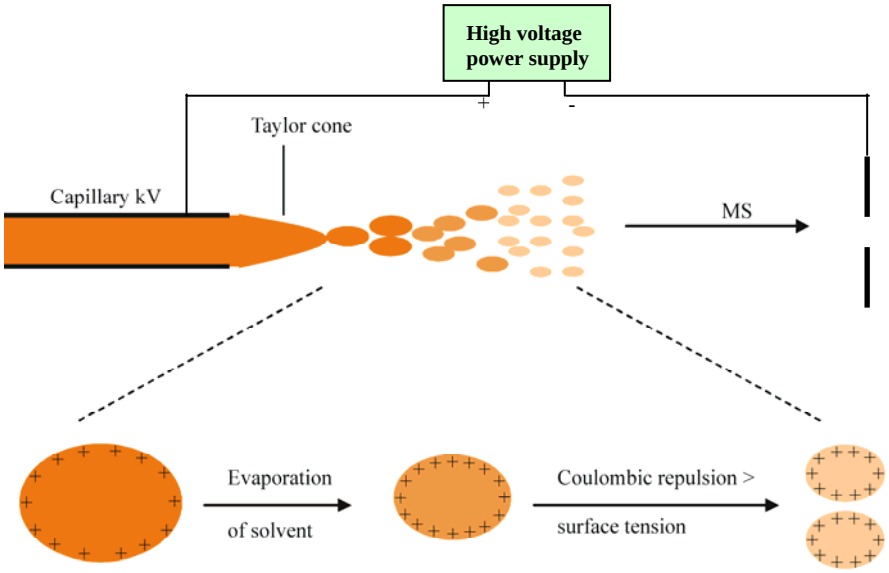


Figure 1.3. The formation of charged droplets and their subsequent fission during ESI.

The characteristic appearance of ESI-spectra of proteins is a sequence of peaks corresponding to heterogeneously multiply charged analytes, each peak differing by one charge from adjacent neighbors in the sequence (Fenn *et al*, 1989). The peaks correspond to the protein mass according to the formula (Mw+n/n), where Mw is the molecular weight of the protein and n is the number of charges. The mass of the protein is deduced from the m/z-spectrum using deconvolution algorithms (Mann *et al*, 1989). Because of the occurrence of multiple charged ions, large analytes can be analyzed with mass analyzers having a limited mass-to-charge range (Mann *et al*, 1995).

Matrix assisted laser desorption/ionization

Kiochi Tanaka accidentally mixed glycerol with a fine metal powder, and the study of analytes dissolved in this mixture eventually resulted in the description of the principles on which MALDI is based (Tanaka *et al*, 1988). The MALDI process however was already described in 1985 when it was found that amino acids and dipeptides mixed with tryptophan can readily be ionized after laser irradiation (Karas *et al*, 1985). Currently, the glycerol-cobalt mixture used by Tanaka is replaced by solid crystals of small acidic molecules containing aromatic moieties. Solutions of analytes mixed with matrix are spotted on metal plates and left to dry, resulting in co-crystallization (Hillenkamp *et al*, 1991).

A pulsed irradiation of these crystals results in energy deposition from the laser onto the matrix molecules, having an adsorption maximum at or near the wavelength of the laser. Matrix molecules and analytes are sublimated from the probe and analyte molecules are

ionized by proton transfer from matrix to analyte (Figure 1.4). The precise nature of the ionization process in MALDI is largely unknown and signal intensities depend on incorporation of analytes into crystals and their likelihood of capturing or retaining a proton during the desorption process. The most recent proposal for MALDI-ionization is the 'lucky survivor model' (Karas *et al*, 2000).

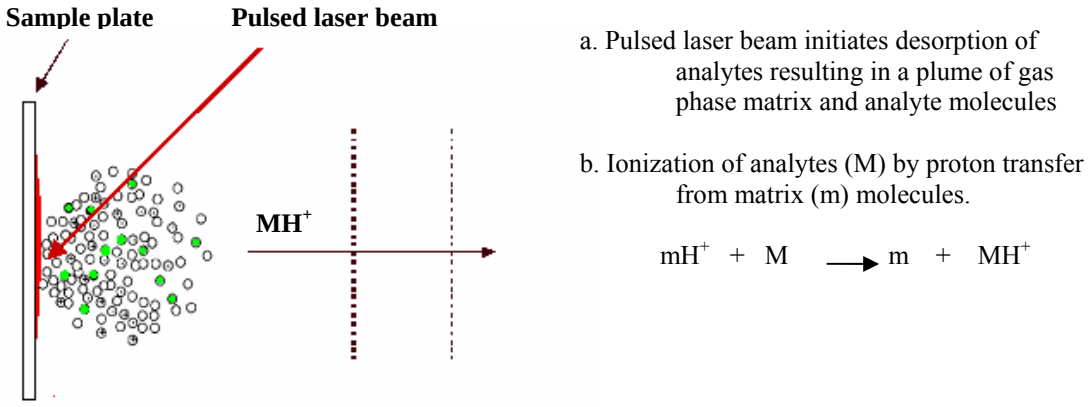


Figure 1.4. MALDI-ionization

After ionization, protonated analytes are extracted from the ion source towards an electrode that is situated in front of the mass analyzer. Different matrices that differ in the amount of energy they impart to the biomolecules were evaluated. For the analysis of certain molecules specific matrixes are preferred such as 4-hydroxy-α-cyanocinnamic acid (CHCA) for peptides and ferulic acid for proteins (Madonna *et al*, 2000). Combination of different matrices, e.g. CHCA and 2,5-dihydroxy benzoic acid (DHB) (Laugesen *et al*, 2003), or the combination of commonly used matrices with salts result in more qualitative spectra. Furthermore, the combinations of matrices and diammonium citrate have been used to enhance signal intensity of acidic peptides (Yang *et al*, 2004a; Oehlers *et al*, 2005; Kinumi *et al*, 2006).

MALDI-ionization is fairly insensitive to salt contamination in the samples. Furthermore, MALDI MS-spectra are easy to interpreter because mainly singly charged ions are generated.

Mass analyzers; Time-Of-Flight (TOF)

Basically, a TOF is a vacuum flight tube with an ion source at one end and a detector at the other (Wiley *et al*, 1955). After ionization, analytes are accelerated towards a counter electrode. After passing through the extraction electrode, the ions enter a field-free drift tube in which they float at a constant speed. At that point, the kinetic energy ($E_{kin}=1/2mv^2$) of an ion in the field-free TOF-tube equals the acceleration energy it has received during the extraction ($E_{acc}=zVe$). From these formulas, it is apparent that the flight time of an ion (t) depends on the mass-to-charge ratio (m/z) of the ion and on the length of the flight tube (x), according to the following formula:

$$t = \left(\frac{m}{z} \frac{x^2}{2Ve} \right)^{1/2}$$

It is evident from this formula that ions with a higher m/z-ratio will reach the detector later than ions with lower m/z-ratios. Because x is constant for a specific instrument, a TOF can be calibrated by the analysis of a single ion with known m/z. Initially, the resolution of TOF-analyzers was limited because not all ions are formed at the same distance from the sample plate and because there is a spread in the velocity that ions initially acquire during desorption.

To improve the resolution of TOF analysis, two technical improvements have been developed. The first one involves the use of an ion mirror at the end of the drift tube (Mamyrin *et al*, 1973). Ions with identical m/z but slightly different velocities are reflected at different points in the mirror and focused on the detector (secondary focusing). The reflectron consists of a series of concentric rings with increasing potential and is maintained at the same potential as the ions. The increase in resolution is due to the refocusing of the ions and because the ions pass through the flight tube twice (increase of factor x in the formula). However, reflectron analysis results in a lower sensitivity and imposes an upper m/z limit (Verentchikov *et al*, 1994). The second method to improve the resolution in TOF-MS is the use of delayed extraction (DE). Instead of applying a continuous extraction field, the extraction is pulsed and applied with a certain delay after the desorption (Brown *et al*, 1995; Vestal *et al*, 1995). DE compensates for the spread in initial velocity of ions with the same m/z. Ions with a lower initial velocity are accelerated more than those having higher initial velocities. Because all ions with the same m/z are focused at the detector at the same time, the resolution of the mass spectra increases (Figure 1.5).

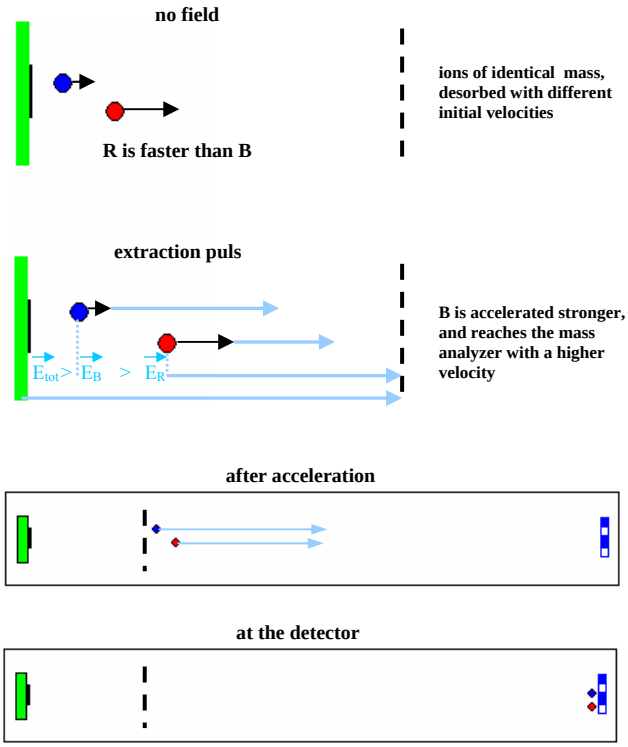


Figure 1.5. Principle of delayed extraction: ions R and B, respectively Red and Blue, with identical m/z are desorbed with different initial velocities. Because ion R, highest initial velocity, has approached the extraction grid more it is accelerated less during the extraction. The result is that both ions are focused at the detector at the same time.

Because of the pulsed nature of both MALDI-ionization and TOF-analysis the combination MALDI-TOF is logic. In 1997 the introduction of the Q-TOF allowed the coupling of ESI to TOF-analysis (Morris *et al*, 1997). Today TOF is perhaps the most sensitive, fastest and robust mass analyzer.

4700 MALDI TOF/TOF

The MALDI-TOF is the instrument of choice for high-throughput peptide mass fingerprint analysis (Part 1.2.2.1). However, despite the development and application of post- and in-source decay (Patterson *et al*, 1994; Kaufmann *et al*, 1996), the MS/MS capabilities are limited because of the poor mass accuracy, the low abundance of product ions and the low

sensitivity of these approaches (Griffin *et al*, 1995; Mann *et al*, 2001). In 2000, a new hybrid instrument (Figure 1.6), having two TOF-flight tubes separated by a collision cell, was described (Medzihradszky *et al*, 2000). The two flight tubes form a continuum in the MS-mode. In the MS/MS mode, ions are initially separated according to their m/z in the first, short, TOF to allow precursor selection. After fragmentation of the selected parent ion in the collision cell, fragment ions generated in the second source are analyzed in the second TOF-tube (Figure 1.7).

A two-gated 'timed-ion-selector' (TIS) at the end of the first flight tube allows high resolution precursor selection (Vestal *et al*, 2005). Because of the high kinetic energy of the ions, both low and high energy collisions can be effected. Therefore, fragment ion spectra of peptides using the 4700 TOF/TOF are typically rich in different ion types, including ions resulting from high-energy side chain fragmentation pathways. Although the presence of ions resulting from side chain fragmentations, the so called d-, v- and w-ions, allows to distinguish the isobaric amino acids Ile and Leu (Vanrobaeys *et al*, 2003), the complexity of the fragmentation spectra hampers *de novo* sequence determination. The CID-cell between the two flight tubes can be operated both at high vacuum or be filled with collision gas. When no gas is present in the collision cell, more specific fragmentation of peptide bonds is observed (Pashkova *et al*, 2005) in a process that resembles metastable decay, but is induced by a difference in potential between the source and the CID-cell. The 4700 TOF/TOF is equipped with a high frequency Nd:YAG laser that operates at a frequency of 200 Hz. This high frequency, together with the high sensitivity inherent to MALDI, has made the 4700 one of the most popular mass spectrometers today allowing fast, sensitive analyses in fully automated settings.

Figure 1.6. The 4700 TOF/TOF

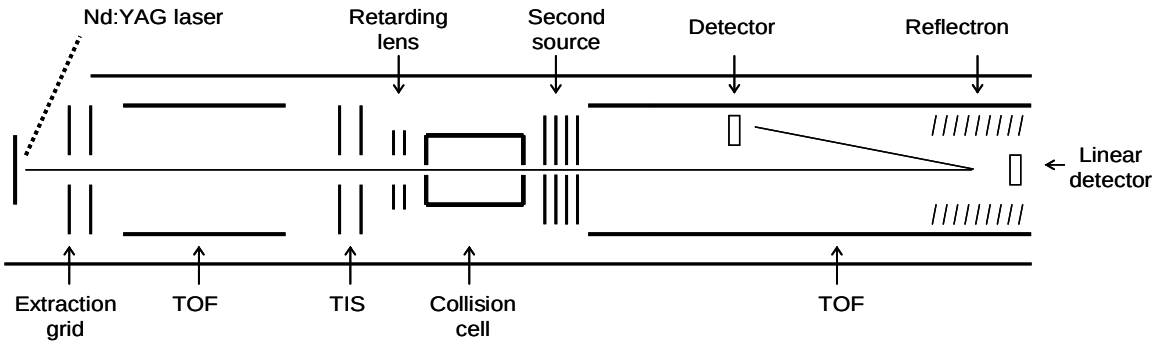


Figure 1.7. Schematic of the 4700 MALDI TOF/TOF, the most important modules are indicated.

1.2.2. Mass spectrometric approaches in proteomics

The use of mass spectrometry for the study of intact proteins, the so called 'top-down' approaches, is discussed in Part 1.2.3. Here, two strategies that are based on the analysis of peptides, 'peptide mass fingerprinting' and 'fragment mass analysis' are depicted. Both strategies rely on the meticulous matching of acquired mass spectra to theoretical spectra obtained by in-silico processing of database entries. Due to its high specificity (Olsen *et al*,

2004) and the favorable ionization properties of the resulting peptides, trypsin digestion is the preferred cleavage method for high-throughput proteomics. For the study of specific properties of proteins (e.g. C-terminal sequence analysis, Part 2), or for the study of specific groups of proteins other cleavage methods are used. Some examples include on-membrane cleavage with CNBr (Kuhn *et al*, 2003) or cleavage with endoprotease LysC (Nielsen *et al*, 2005) for the study of membrane proteins. Alternatively, more than one cleavage method, with different specificity, have been applied to attain complete sequence coverage (Choudhary *et al*, 2003; Cravello *et al*, 2003; Wu *et al*, 2003). A third MS-based proteome analysis is ladder sequence analysis (Part 2). Here, concatenated sets of N- or C-terminal sequentially truncated peptides are generated and subsequently analyzed to obtain sequence information (Chait *et al*, 1993; Samyn *et al*, 2005). Recently, a high-throughput ladder sequence analysis method was described. In this approach microwave-assisted controlled partial acid hydrolysis of proteins is used to generate sequence ladders from the N- and the C-terminus of proteins (Zhong *et al*, 2004).

1.2.2.1. Peptide mass fingerprinting

The simplest and fastest strategy for MS-based protein identification involves the proteolytic generation of peptides and the comparison of the resulting 'peptide mass map' with theoretical mass maps. Because resemblance between the experimental mass map and the theoretical map must exist, the use of this strategy is limited to the analysis of isolated proteins or simple mixtures. Consequently, it can only be used for the identification of proteins separated with 2D-PAGE or liquid chromatography.

Peptide mass mapping is based on the knowledge that the accurate mass of a group of peptides derived from a protein by sequence-specific proteolysis is a highly effective means of protein identification. Therefore, the concept behind peptide mass fingerprint analysis is quite simple and was independently implemented by several groups at approximately the same time (Henzel *et al*, 1993; James *et al*, 1993; Mann *et al*, 1993; Pappin *et al*, 1993; Yates *et al*, 1993). Different proteins will, after proteolysis with a specific protease, produce groups of peptides, the masses of which constitute mass fingerprints unique for a specific protein (Figure 1.8). Therefore, when the peptide mass fingerprint is used as query for searching a sequence database containing the protein sequence, the protein is expected to be correctly identified within the database (Aebersold *et al*, 2001).

Factors that influence the result of commonly used algorithms for PMF include mass accuracy, spectrum quality and sequence coverage. Increased mass accuracy of the analysis will decrease the number of isobaric peptides for any given mass in a sequence database and, therefore, the stringency of the search increases. When mass accuracies of 1 ppm are reached, a single peptide mass, a so called 'accurate mass tag', should be sufficient to identify protein entries in small databases (Clauser *et al*, 1999).

In general, only a part of the predicted peptides are observed in peptide mass maps. This is not considered to be a problem since only a relatively low number of peptide masses is required for protein identification. Unassigned peptide masses, on the contrary, are a significant problem (Thiede *et al*, 2005). These may have different origins such as changes in the expected peptide masses by posttranslational modifications and artifactual modifications due to sample handling. Furthermore, sequence variants and the presence of errors in databases can result in low significant identifications and false positives (Karty *et al*, 2002). Allowing numerous modifications in the search and widening of the tolerated mass fault will

lower the number of unassigned peaks but increases the likelihood of false positive identification.

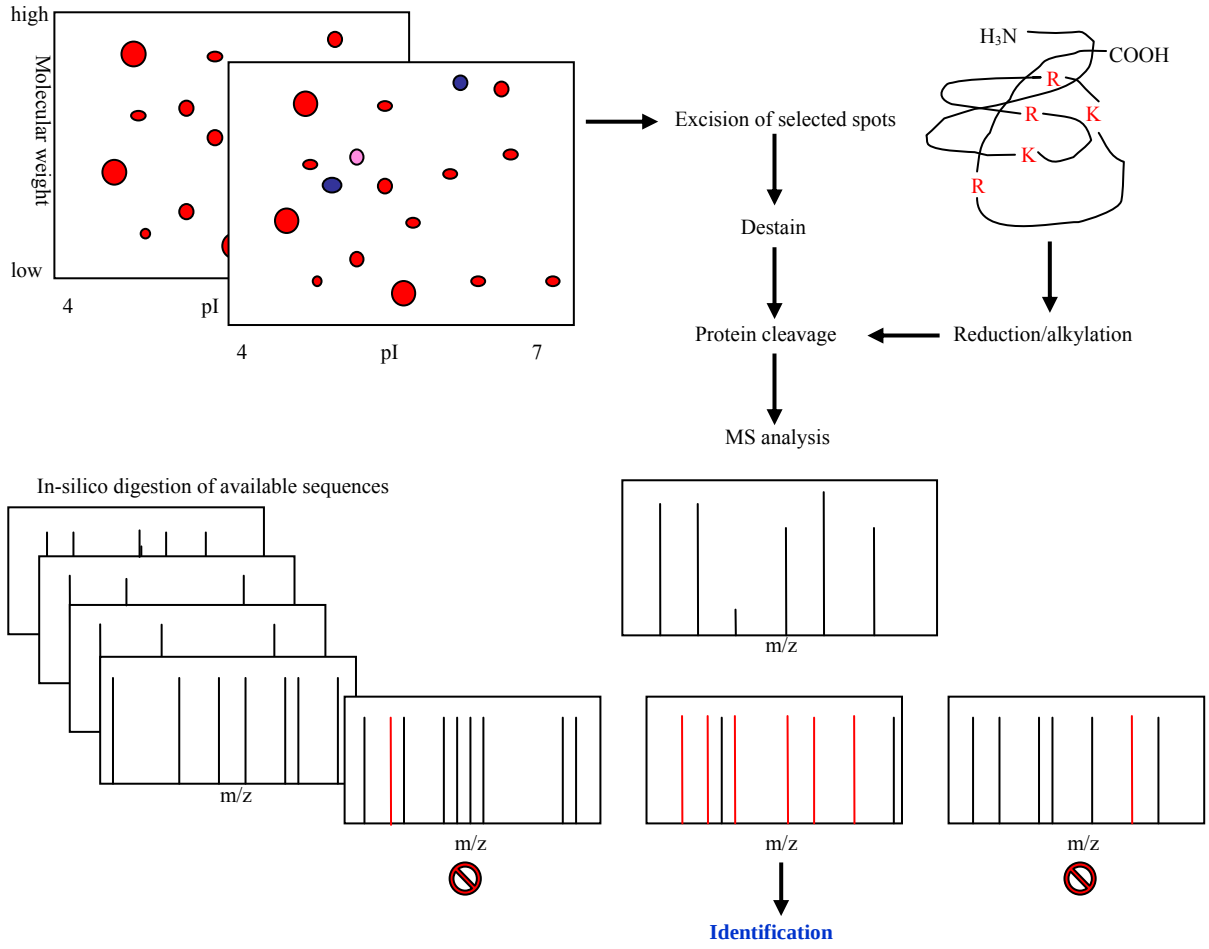


Figure 1.8. Peptide mass fingerprinting. Proteins isolated using 2D-PAGE or chromatographic approaches are digested and a mass spectrum of the resulting peptides is acquired. Matching of the experimentally observed masses with theoretical mass maps derived from database entries, applying user defined settings, results in identification of the protein.

Iterative cycles of PMF interpretation allow a more in-depth analysis of PMF-spectra. This can be used to remove peptides associated with an unambiguous match from the query (Jensen *et al*, 1997), or to remove contaminating masses (Schmidt *et al*, 2003). Despite the robustness of PMF, its reliability as sole tool for protein identification is currently a matter of debate and confirmation of identifications using MS/MS is often required (Carr *et al*, 2004).

1.2.2.2. Fragment mass analysis

Different amino acid compositions and permutations of an amino acid can result in peptides with identical mass. Therefore, the amino acid sequence of a peptide is more constraining for protein identification by sequence database searching than its mass (Zubarev *et al*, 1996). The limited amount of information that is acquired during peptide mass mapping can thus be supplemented with structural information obtained in MS/MS experiments. *De novo* sequence determination after fragmentation of peptides is more elaborately discussed in Part 3.1. Here, general considerations and automated identification strategies are discussed.

As mentioned above, tandem mass spectrometry requires the use of two mass analyzers, the first for selection of the precursor-peptide of interest and the second for analysis of the fragments. Only peptides selected in the first mass analyzer reach the collision cell. In the cell, analytes are excited to higher energies through collisions with neutral gas molecules. This excess energy is dissipated through fragmentation of the peptides in a process called 'collision induced dissociation' (CID) or synonymously 'collision actived dissociation' (CAD). For the designation of the resulting fragments, a normalized nomenclature has been proposed (Figure 1.9) (Roepstorff *et al*, 1984; Johnson *et al*, 1987).

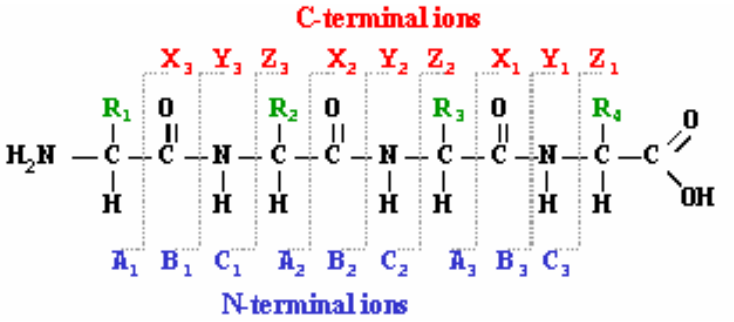


Figure 1.9. Nomenclature of the backbone fragment ions of peptides observed during tandem mass spectrometry.

The fragmentation mechanisms and the preferences for fragmentation of specific peptide bonds have recently been reviewed (Paizs *et al*, 2005). Here, the approaches depicted in the central and left part of Figure 1.10, non-error tolerant database searches, are discussed. These approaches are based on spectral comparison of experimental data with in-silico constructed fragmentation spectra of peptides, so called 'peptide fragmentation fingerprinting' analysis (PFF). Alternatively, the use of short sequences together with the limiting masses, the so called 'sequence tags', can be used for database searches.

A first approach for computer-assisted protein identification via CID spectra uses the uninterpreted fragment ion pattern and the mass of the parent ion as input for sequence database searching. Several search algorithms for PFF-analysis have been developed, the most generally used being Mascot and SEQUEST (Eng *et al*, 1994; Perkins *et al*, 1999). These algorithms first create a list of peptide masses which are isobaric to the selected precursor. The algorithms next allow to calculate the masses of the expected fragment ions from each of the peptides on this list. Chemical information that would be reflected in the relative intensity of the fragment ions, such as the influence of the sequence on the fragmentation, is generally not taken into account. Subsequently, these theoretical fragmentation spectra are compared with the observed fragment ion spectrum and a score reflecting the similarity between the spectra is calculated. Because of the stringent matching of theoretical and experimental spectra, these algorithms can only be successfully applied when the peptide sequence of the fragmented precursor is represented in the database. Furthermore, because knowledge of fragmentation mechanisms is not used in the prediction of fragmentation spectra, manual validation of the predicted sequence is often required to avoid false positive identifications. Nevertheless, the constraints on database searching using a fragmentation spectrum are so stringent that the fragmentation spectrum of a single peptide can be adequate for protein identification, even when only 'expressed sequence tag'-databases (EST) are available for the organism under study. The approach is easily automated, and expected posttranslational modifications can be identified by indicating the programs to anticipate mass shift at specific residues. Despite the power of these programs, concern was expressed regarding the fact that

only a fraction of the tandem mass spectra in large-scale proteome projects are effectively used in protein identification (Peng *et al*, 2003; Resing *et al*, 2004).

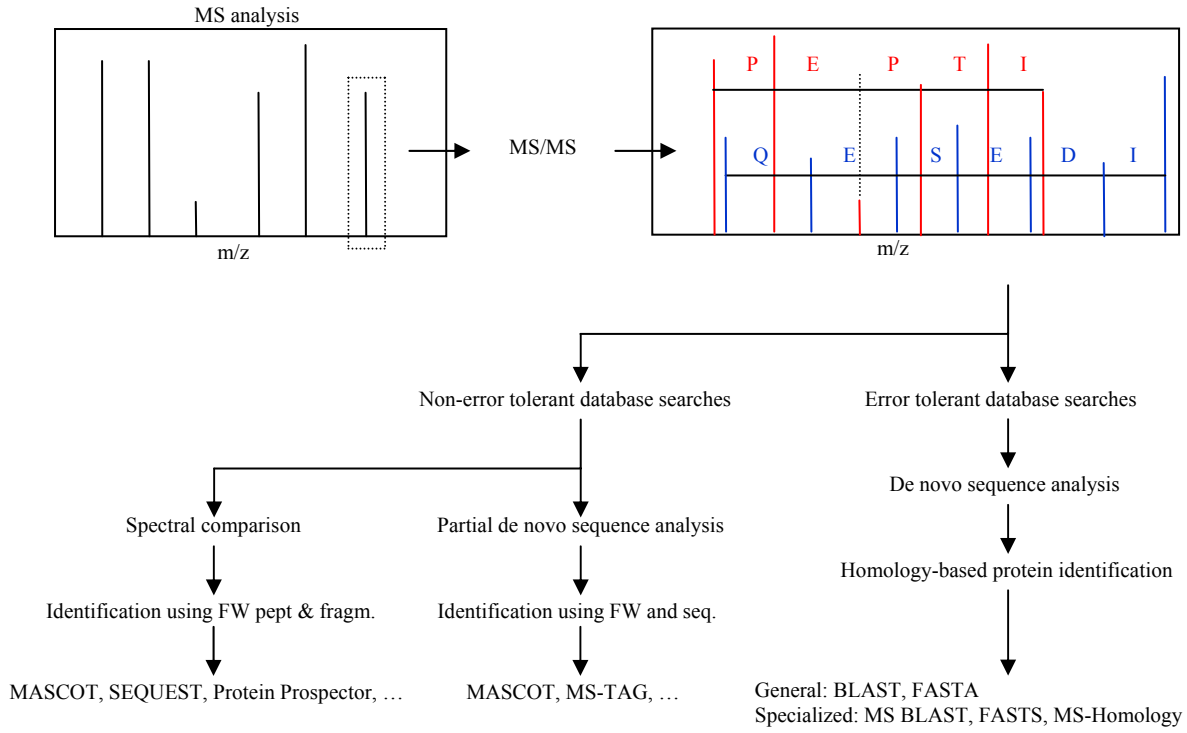


Figure 1.10. Peptide fragmentation analysis. The different routes to identify proteins based on peptide fragmentation spectra are depicted.

Sequence tag-approaches rely on the manual or automated interpretation of a part of the fragmentation spectra. In this way, consecutive elements of a particular ion series that provides a partial sequence are identified. Although the determination of complete peptide sequences from MS/MS spectra is often prohibitively challenging, incomplete peptide sequences are often readily determined (Mann *et al*, 1994). This partial sequence is then used in database searches, together with the mass of the parent ion, the mass where the determined sequence starts and the mass difference between the end of the tag and the parent mass (Figure 1.11). A similar approach is used for protein identification after fragmentation of intact proteins (Mortz *et al*, 1996). Database searches, combining the sequence tags generated from all the peptides present in a single spot have been used for cross-species protein identification using the MultiTag-algorithm (Sunyaev *et al*, 2003).

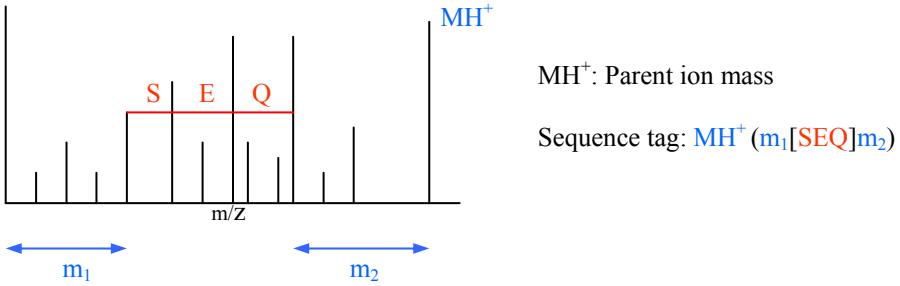


Figure 1.11. The peptide sequence tag approach

1.2.3. Analysis of intact proteins

Compared to the other 'top-down' analysis method, separation of intact proteins with gel electrophoresis, MS analysis of intact proteins offers a higher resolution (Kelleher, 2004). Top-down MS analysis of proteins offers the possibility of interrogation of the complete DNA-predicted sequence. Efficient surveying of an entire protein sequence with 100% coverage at the intact and fragment ion level (e.g. a complementary pair) is the hallmark of a top-down MS experiment (Forbes *et al*, 2001). Conceptually, the analysis of the mass of a protein is the simplest way to determine the identity of the protein and to determine the presence of posttranslational modifications and/or proteolytic processing (Horn *et al*, 2000). For the unambiguous localization of the exact site of PTMs, fragment ion analysis of proteins is required (Figure 1.12). This was initially demonstrated using triple-quadrupole instruments but is currently mainly associated with Fourier transform mass spectrometry (FTMS) analysis (Loo *et al*, 1990; Loo *et al*, 1992).

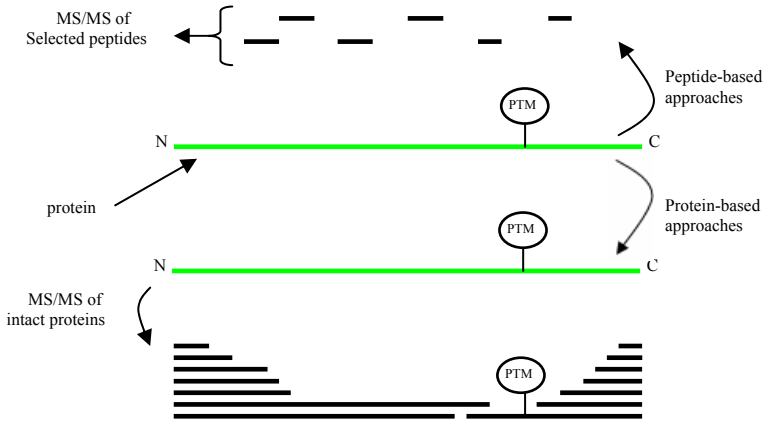


Figure 1.12. Bottom-up versus top-down MS approaches. Sequence coverage is typically between 7-70% for bottom-up and 100% for top-down approaches.

Top-down analysis has been performed using a variety of mass spectrometers including triple-quadrupoles and MALDI-TOF/TOF (Lin *et al*, 2003; Suckau *et al*, 2003). Compared to other mass spectrometers, FTMS instruments are expensive and experienced operators are not commonly in most research groups. However, only FTMS provides sufficient resolution to resolve isotopic peaks, this in combination with a high mass accuracy. Due to recent developments, the limitations in top-down MS analysis are no longer situated in the mass analysis but rather in the front-end sample handling (Meng *et al*, 2002), data processing and computer-aided interpretation (Kelleher, 2004). More efficient fragmentation mechanisms (Yamada *et al*, 2006) and an initial study on relative quantification in top-down approaches (Du *et al*, 2006) further demonstrate the potential of this approach..

1.2.4. Quantification in proteomics

As described before, the acquisition of gene expression profiles using microarrays provides a fast and sensitive tool to quantitatively determine changes in transcription. Nevertheless, because of the limited correlation between mRNA and protein abundance, quantification of all proteins in a cellular extract is necessary to determine the abundance of proteins (Gygi *et al*, 1999b). Furthermore, because non-stoichiometric modifications can have an important functional impact, the quantity of each protein form should be determined in order to describe the functional properties of a biological system. Quantification of proteins can be performed either based on the intensity of spots on gel images or by using derivatization strategies with isotopic labels (Righetti *et al*, 2004a).

Quantification of protein expression using 2D-separated proteins has recently been reviewed (Righetti *et al*, 2004b) and is based on the processing of gel images. Using

appropriate software packages, comparison of the staining intensity on gel images from different cellular states, e.g. healthy versus diseased tissue, is used for relative quantification. However, the use of staining procedures with dyes such as 'Coomassie Brilliant Blue' and silver requires that multiple gels from each sample are run. The introduction of fluorescent dyes only improved the sensitivity and the dynamic range of such comparisons (Rabilloud, 2000). Another approach to compare the protein abundance in specific cell states is the DIGE-method (**D**ifferential **I**n-**G**el **E**lectrophoresis) (Unlu *et al*, 1997). Samples from differing cellular states are labeled with fluorescent dyes with a different excitation and emission maximum. The samples are pooled and run in a single gel analysis; the spots are visualized at different wavelengths with specialized scanners, and software is used to analyze the gels (Zhou *et al*, 2002). Running the samples in a single gel eliminates gel-to-gel variation and the use of fluorescent dyes results in a high sensitivity. The sensitivity of the DIGE-approach was further increased by using similar dyes for maximal labeling at cysteine residues (Shaw *et al*, 2003). This saturation labeling results in a high detection sensitivity but has detrimental effects on the subsequent mass spectrometric identification of proteins (Kondo *et al*, 2003).

In bottom-up approaches, relative quantification of proteins in different samples is performed using stable isotope tagging and mass spectrometry (Tao *et al*, 2003). Mass spectra, MS or MS/MS, of the mixed differentially labeled samples reveal characteristic mass shifts between peptides that are chemically identical but mass-differentiated. The relative intensity of the corresponding peaks in mass spectra is a measure for their relative abundance. The prototype of this quantification method is ICAT. In this approach, cysteines in the two samples are reacted with an isotopically labeled biotin-containing affinity tag (Gygi *et al*, 1999a). After labeling, the samples are mixed, digested and cysteine-containing peptides are isolated with avidin-affinity chromatography. The use of the original ICAT reagent has several drawbacks, related to the bulkiness of the derivative and the chromatographic separation of differentially labeled peptides. Therefore, new reagents have been developed in which a cleavable linker is introduced as well as the use of ^{13}C instead of deuterium (Williamson *et al*, 2002). Other quantitative methods that use mass shifts in MS-analysis for quantification include '**S**table **I**sotope **L**abeling by **A**mino acids in **C**ulture' (SILAC) (Ong *et al*, 2002) and proteolytic ^{18}O-labeling (Yao *et al*, 2001).

Another strategy for relative quantification, isobaric tagging, was introduced with a method called '**i**sobaric **T**agging for **R**elative and **A**bsolute **Q**uantification' (iTRAQ) (Ross *et al*, 2004). Using ICAT or comparable methods, the peptide signal in MS-mode is divided over different peaks; this potential loss in sensitivity is avoided in isobaric tagging strategies. The iTRAQ reagent is amine reactive with a fixed mass, containing a reporter group and a group that balances the mass differences in the reporter groups (Figure 1.13). During CID-analysis the bond between the reporter- and the balance-group readily breaks and the reporter is observed in MS/MS spectra at 114, 115, 116 or 117 Da (Aggarwal *et al*, 2005). Using iTRAQ, the relative quantity of a protein in four samples can be determined in a single experiment; i-PROT even allows multiplexing of up to six samples. In this programme, the used isobaric label is an oligopeptide, GGGGGGDPGGGGGG, having an N-terminal reactive group attached. The reactivity of this group can be varied and determines the peptide subset that is selected, i.e. cysteine-containing, lysine-containing or phosphoserine-containing peptides. CID of tagged peptides results in the fragmentation of the DP-bond and relative quantification can be done using both low (differentially labeled y-like ion PGGGGGG) and high mass signals (Latimer *et al*, 2004).

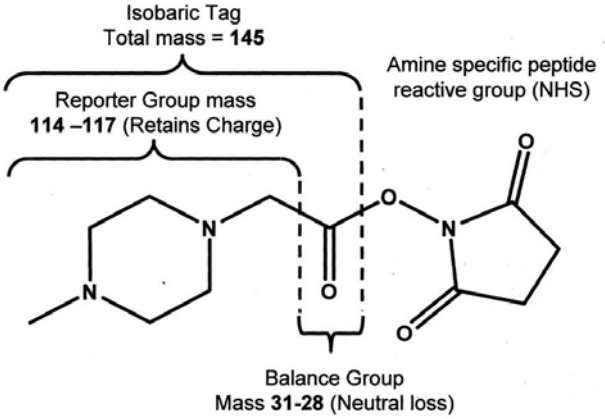


Figure 1.13. Structure of the iTRAQ-reagent

All approaches described above were recently compared and found to perform equally well for relative quantitative proteome analysis (Wu *et al*, 2006). A method for absolute quantification was developed and coined AQUA. Instead of using two similar samples, peptide mixtures are spiked with a known amount of an isotopically labeled synthetic peptide. The intensity ratio of the synthetic versus the target peptide allows absolute quantification of the target peptide/protein in the sample (Gerber *et al*, 2003).

1.2.5. Posttranslational modifications

More than 300 different protein posttranslational modifications (PTMs) were known in 1993 and new ones are regularly described (Krishna *et al*, 1993). These include diverse processes as proteolysis, controlled protein degradation, phosphorylation, lipidation, S-nitrosylation, nitration, oxidation, glycosylation, methylation, adenosine diphosphate (ADP)-ribosylation, acylation, ubiquitination, sulfation and farnesylation. PTMs change the size, charge, structure and conformation of proteins. As a result, characteristics of proteins, such as enzyme activity, binding affinity, and protein hydrophobicity, are altered. PTMs cannot only directly change the protein function, but also indirectly affect function by leading to cell compartmentalization, sequestration, degradation, elimination, and protein–protein interaction.

Despite the great importance of PTMs for biological function, their large-scale study has been hampered by a lack of suitable methods, and many key modifications have only been discovered late in the elucidation of biological processes. As a result, the extent and functional importance of protein modifications on the activities of the cell are probably not yet fully realized (Mann *et al*, 2003; Cantin *et al*, 2004). Although some modifications can be predicted from consensus sequences or by the analysis of the activity of the proteins that modify other proteins (Blom *et al*, 2004; Kiemer *et al*, 2005), the amount of PTMs is generally overestimated using this approach and dynamic and temporal properties of posttranslational modifications are not considered (Jensen, 2000). Many PTMs have been discovered serendipitously during studies of individual proteins with the help of standard molecular techniques, such as deletion of the amino acids bearing the modification. Direct analysis of modifications requires isolation of the processed protein in a sufficiently large amount for (bio)chemical study. Nevertheless, general, high-throughput, approaches for protein identification can also result in the identification of PTMs. For instance, several PTMs were characterized in a study where multiple digests were performed and the samples exhaustively analyzed by MDLC (MacCoss *et al*, 2002).

Methods specific for the analysis of PTMs were mainly developed for the analysis of phosphorylation and glycosylation. The study of glycosylation is particularly challenging because carbohydrate moieties typically have a branched, complex structure. Furthermore, different monosaccharides have the same mass, which hampers unambiguous structural determinations using mass spectrometry (Reinhold *et al*, 1995). Different methods have been developed to specifically 'fish' carbohydrate carrying proteins out of complex samples (Hirabayashi *et al*, 2002). Lectin-glycoprotein interactions are the main tools used for the isolation of glycoproteins from cellular protein extracts, affording selectivity on the type of carbohydrate that is isolated by using lectins of differing specificity (Yang *et al*, 2004b).

For the analysis of phosphorylated proteins, methods based on affinity-enrichment, immuno-recognition, chemical derivatization and radioactive labeling have been described and recently reviewed (Reinders *et al*, 2005). This multitude of methods reflects the importance of this type of PTM. Phosphorylation events act as a major regulator of signaling processes and as a reversible on-off switch in metabolic pathways. The fact that protein kinases are one of largest enzyme families known further illustrates their biological importance (Hunter, 2000).

References

Aebersold, R.; Goodlett, D.R. (**2001**). Mass spectrometry in proteomics. *Chem Rev* **101**(2): 269-95.

Aggarwal, K.; Choe, L.H.; Lee, K.H. (**2005**). Quantitative analysis of protein expression using amine-specific isobaric tags in Escherichia coli cells expressing rhsA elements. *Proteomics* **5**(9): 2297-308.

Alaiya, A.; Al-Mohanna, M.; Linder, S. (**2005**). Clinical cancer proteomics: promises and pitfalls. *J Proteome Res* **4**(4): 1213-22.

Baggerman, G.; Verleyen, P.; Clynen, E.*; et al.* (**2004**). Peptidomics. *J Chromatogr B* **803**(1): 3-16.

Barry, R.; Diggle, T.; Terrett, J.*; et al.* (**2003**). Competitive assay formats for high-throughput affinity arrays. *J Biomol Screen* **8**(3): 257-63.

Barry, R.; Soloviev, M. (**2004**). Quantitative protein profiling using antibody arrays. *Proteomics* **4**(12): 3717-26.

Bateman, A.; Birney, E.; Cerruti, L.*; et al.* (**2002**). The Pfam protein families database. *Nucleic Acids Res* **30**(1): 276-80.

Bauer, A.; Kuster, B. (**2003**). Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. *Eur J Biochem* **270**(4): 570-8.

Bertone, P.; Snyder, M. (**2005**). Prospects and challenges in proteomics. *Plant Physiol* **138**(2): 560-2.

Blom, N.; Sicheritz-Ponten, T.; Gupta, R.*; et al.* (**2004**). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**(6): 1633-49.

Borodovsky, M.; Rudd, K.E.; Koonin, E.V. (**1994**). Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res* **22**(22): 4756-67.

Brown, R.S.; Lennon, J.J. (**1995**). Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption ionization linear time-of-flight mass spectrometer. *Anal Chem* **67**(13): 1998-2003.

Bunai, K.; Yamane, K. (**2005**). Effectiveness and limitation of two-dimensional gel electrophoresis in bacterial membrane protein proteomics and perspectives. *J Chromatogr B* **815**(1-2): 227-36.

Cantin, G.T.; Yates, J.R., 3rd (**2004**). Strategies for shotgun identification of post-translational modifications by mass spectrometry. *J Chromatogr A* **1053**(1-2): 7-14.

Carpentier, S.C.; Witters, E.; Laukens, K.*; et al.* (**2005**). Preparation of protein extracts from recalcitrant plant tissues: an evaluation of different methods for two-dimensional gel electrophoresis analysis. *Proteomics* **5**(10): 2497-507.

Carr, S.; Aebersold, R.; Baldwin, M.*; et al.* (**2004**). The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* **3**(6): 531-3.

Cekan, S.Z. (**2004**). Methods to find out the expression of activated genes. *Reprod Biol Endocrinol* **2**(1): 68.

Chait, B.T.; Wang, R.; Beavis, R.C.*; et al.* (**1993**). Protein ladder sequencing. *Science* **262**(5130): 89-92.

Choudhary, G.; Wu, S.L.; Shieh, P.*; et al.* (**2003**). Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *J Proteome Res* **2**(1): 59-67.

Clauser, K.R.; Baker, P.; Burlingame, A.L. (**1999**). Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* **71**(14): 2871-82.

Cravello, L.; Lascoux, D.; Forest, E. (**2003**). Use of different proteases working in acidic conditions to improve sequence coverage and resolution in hydrogen/deuterium exchange of large proteins. *Rapid Commun Mass Spectrom* **17**(21): 2387-93.

Cusick, M.E.; Klitgord, N.; Vidal, M.*; et al.* (**2005**). Interactome: gateway into systems biology. *Hum Mol Genet* **14** (Spec No. 2): R171-81.

Dole, M.; Mack, L.L.; Hines, R.L.*; et al.* (**1968**). Molecular beams of macroions. *J Chem Phy* **49**: 2240-9.

Du, Y.; Parks, B.A.; Sohn, S.*; et al.* (**2006**). Top-down approaches for measuring expression ratios of intact yeast proteins using fourier transform mass spectrometry. *Anal Chem* **78**(3): 686-94.

Duan, X.; Yarmush, D.; Berthiaume, F.*; et al.* (**2005**). Immunodepletion of albumin for two-dimensional gel detection of new mouse acute-phase protein and other plasma proteins. *Proteomics* **5**(15): 3991-4000.

Eisen, M.B.; Spellman, P.T.; Brown, P.O.*; et al.* (**1998**). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**(25): 14863-8.

Eng, J.K.; McCormack, A.L.; Yates, J.R.I. (**1994**). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**(11): 976-89.

Farhoud, M.H.; Wessels, H.J.; Wevers, R.A.*; et al.* (**2005**). Serial isoelectric focusing as an effective and economic way to obtain maximal resolution and high-throughput in 2D-based comparative proteomics of scarce samples: proof-of-principle. *J Proteome Res* **4**(6): 2364-8.

Fenn, J.B.; Mann, M.; Meng, C.K.*; et al.* (**1989**). Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**(4926): 64-71.

Fenn, J.B. (**1993**). Ion formation from charged droplets: roles of geometry, energy and time. *J Am Soc Mass Spectrom* **4**(7): 524-35.

Fiehn, O. (**2002**). Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* **48**(1-2): 155-71.

Forbes, A.J.; Mazur, M.T.; Patel, H.M.*; et al.* (**2001**). Toward efficient analysis of >70 kDa proteins with 100% sequence coverage. *Proteomics* **1**(8): 927-33.

Fridman, E.; Pichersky, E. (**2005**). Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. *Curr Opin Plant Biol* **8**(3): 242-8.

Fukusaki, E.; Kobayashi, A. (**2005**). Plant metabolomics: potential for practical operation. *J Biosci Bioeng* **100**(4): 347-54.

Garbis, S.; Lubec, G.; Fountoulakis, M. (**2005**). Limitations of current proteomics technologies. *J Chromatogr A* **1077**(1): 1-18.

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

Gevaert, K.; Goethals, M.; Martens, L.*; et al.* (**2003**). Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol* **21**(5): 566-9.

Gevaert, K.; Van Damme, P.; Martens, L.*; et al.* (**2005**). Diagonal reverse-phase chromatography applications in peptide-centric proteomics: ahead of catalogue-omics? *Anal Biochem* **345**(1): 18-29.

Gilar, M.; Olivova, P.; Daly, A.E.*; et al.* (**2005**). Orthogonality of separation in two-dimensional liquid chromatography. *Anal Chem* **77**(19): 6426-34.

Gorg, A.; Obermaier, C.; Boguth, G.*; et al.* (**2000**). The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21**(6): 1037-53.

Gorg, A.; Weiss, W.; Dunn, M.J. (**2004**). Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **4**(12): 3665-85.

Granger, J.; Siddiqui, J.; Copeland, S.*; et al.* (**2005**). Albumin depletion of human plasma also removes low abundance proteins including the cytokines. *Proteomics* **5**(18): 4713-8.

Griffin, P.R.; MacCoss, M.J.; Eng, J.K.*; et al.* (**1995**). Direct database searching with MALDI-PSD spectra of peptides. *Rapid Commun Mass Spectrom* **9**(15): 1546-51.

Griffiths, W.J. **(2000)**. Nanospray mass spectrometry in protein and peptide chemistry. ed. Jolles, P. Jornvall, H. *Proteomics in functional genomics*. Basel, Birkhauser Verlag. **88:** 69-79.

Gygi, S.P.; Rist, B.; Gerber, S.A.*; et al.* (**1999a**). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**(10): 994-9.

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

Haab, B.B.; Dunham, M.J.; Brown, P.O. (**2001**). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol* **2**(2): RESEARCH0004.

Harder, A.; Wildgruber, R.; Nawrocki, A.*; et al.* (**1999**). Comparison of yeast cell protein solubilization procedures for two-dimensional electrophoresis. *Electrophoresis* **20**(4-5): 826-9.

Hattan, S.J.; Marchese, J.; Khainovski, N.*; et al.* (**2005**). Comparative study of [Three] LC-MALDI workflows for the analysis of complex proteomic samples. *J Proteome Res* **4**(6): 1931-41.

Hecker, M. (**2005**). Towards a comprehensive understanding of bacterial physiology by proteomics. *HUPO 4th Annual World Congress, Munich. Mol. Cell. Proteomics* **4**(8): S7.

Henzel, W.J.; Billeci, T.M.; Stults, J.T.*; et al.* (**1993**). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A* **90**(11): 5011-5.

Hild, M.; Beckmann, B.; Haas, S.A.*; et al.* (**2003**). An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. *Genome Biol* **5**(1): R3.

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

Hirabayashi, J.; Kasai, K. (**2002**). Separation technologies for glycomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **771**(1-2): 67-87.

Horn, D.M.; Zubarev, R.A.; McLafferty, F.W. (**2000**). Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc Natl Acad Sci U S A* **97**(19): 10313-7.

Hunter, T. (**2000**). Signaling--2000 and beyond. *Cell* **100**(1): 113-27.

Iribarne, J.V.; Thomson, B.A. (**1976**). On the evaporation of small ions from charged droplets. *J Chem Phy* **64**: 2287-94.

Issaq, H.J.; Chan, K.C.; Janini, G.M.*; et al.* (**2005**). Multidimensional separation of peptides for effective proteomic analysis. *J Chromatogr B* **817**(1): 35-47.

James, P.; Quadroni, M.; Carafoli, E.*; et al.* (**1993**). Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun* **195**(1): 58-64.

Jensen, O.N.; Podtelejnikov, A.V.; Mann, M. (**1997**). Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal Chem* **69**(23): 4741-50.

Jensen, O.N. (**2000**). Modification-specific proteomics: systematic strategies for analysing post-translationally modified proteins. *Trends Biotechnol* **18**(S1): 36-42.

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

Jung, E.; Heller, M.; Sanchez, J.C.*; et al.* (**2000**). Proteomics meets cell biology: the establishment of subcellular proteomes. *Electrophoresis* **21**(16): 3369-77.

Juraschek, R.; Dulcks, T.; Karas, M. (**1999**). Nanoelectrospray--more than just a minimized-flow electrospray ionization source. *J Am Soc Mass Spectrom* **10**(4): 300-8.

Kang, X.; Bates, R.C.; Frey, D.D. (**2000**). High-performance chromatofocusing using linear and concave pH gradients formed with simple buffer mixtures. II. Separation of proteins. *J Chromatogr A* **890**(1): 37-43.

Kang, X.; Frey, D.D. (**2004**). Chromatofocusing of peptides and proteins using linear pH gradients formed on strong ion-exchange adsorbents. *Biotechnol Bioeng* **87**(3): 376-87.

Karas, M.; Bachmann, D.; Hillenkamp, F. (**1985**). Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Anal Chem* **57**(14): 2935-9.

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): 1-12.

Karty, J.A.; Ireland, M.M.; Brun, Y.V.*; et al.* (**2002**). Artifacts and unassigned masses encountered in peptide mass mapping. *J Chromatogr B Analyt Technol Biomed Life Sci* **782**(1-2): 363-83.

Kaufmann, R.; Chaurand, P.; Kirsch, D.*; et al.* (**1996**). Post-source decay and delayed extraction in matrix-assisted laser desorption/ionization-reflectron time-of-flight mass spectrometry. Are there trade-offs? *Rapid Commun Mass Spectrom* **10**(10): 1199-208.

Kebarle, P.; Tang, L. (**1993**). From ions in solution to ions in the gas phase: the mechanism of electrospray mass spectrometry. *Anal Chem* **65**: 972A-986A.

Kebarle, P.; Ho, Y. **(1997)**. On the mechanism of electrospray mass spectrometry. ed. Cole, R. B. *Electrospray ionization mass spectrometry: fundamentals, instrumentation and applications*. New York, Wiley**:** 3-63.

Kebarle, P. (**2000**). A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *J Mass Spectrom* **35**(7): 804-17.

Kelleher, N.L. (**2004**). Top-down proteomics. *Anal Chem* **76**(11): 197A-203A.

Kiemer, L.; Bendtsen, J.D.; Blom, N. (**2005**). NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics* **21**(7): 1269-70.

Kinumi, T.; Shimomae, Y.; Arakawa, R.*; et al.* (**2006**). Effective detection of peptides containing cysteine sulfonic acid using matrix-assisted laser desorption/ionization and laser desorption/ionization on porous silicon mass spectrometry. *J Mass Spectrom* **41**(1): 103-12.

Kondo, T.; Seike, M.; Mori, Y.*; et al.* (**2003**). Application of sensitive fluorescent dyes in linkage of laser microdissection and two-dimensional gel electrophoresis as a cancer proteomic study tool. *Proteomics* **3**(9): 1758-66.

Krishna, R.G.; Wold, F. (**1993**). Post-translational modification of proteins. *Adv Enzymol Relat Areas Mol Biol* **67**: 265-98.

Kuhn, K.; Thompson, A.; Prinz, T.*; et al.* (**2003**). Isolation of N-terminal protein sequence tags from cyanogen bromide cleaved proteins as a novel approach to investigate hydrophobic proteins. *J Proteome Res* **2**(6): 598-609.

Latimer, D.R.; Guerra, C.E.; Feng, L.*; et al.* (**2004**). Biomarker discovery and profiling using highly multiplexed i-PROT labels. *Proceedings of the 52th Conference on Mass Spectrometry and Allied Topics, Nashville, TN*.

Laugesen, S.; Roepstorff, P. (**2003**). Combination of two matrices results in improved performance of MALDI MS for peptide mass mapping and protein analysis. *J Am Soc Mass Spectrom* **14**(9): 992-1002.

Lin, M.; Campbell, J.M.; Mueller, D.R.*; et al.* (**2003**). Intact protein analysis by matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **17**(16): 1809-14.

Loo, J.A.; Edmonds, C.G.; Smith, R.D. (**1990**). Primary sequence information from intact proteins by electrospray ionization tandem mass spectrometry. *Science* **248**(4952): 201-4.

Loo, J.A.; Quinn, J.P.; Ryu, S.I.*; et al.* (**1992**). High-resolution tandem mass spectrometry of large biomolecules. *Proc Natl Acad Sci U S A* **89**(1): 286-9.

Luche, S.; Santoni, V.; Rabilloud, T. (**2003**). Evaluation of nonionic and zwitterionic detergents as membrane protein solubilizers in two-dimensional electrophoresis. *Proteomics* **3**(3): 249-53.

MacCoss, M.J.; McDonald, W.H.; Saraf, A.*; et al.* (**2002**). Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc Natl Acad Sci U S A* **99**(12): 7900-5.

Madera, M.; Mechref, Y.; Novotny, M.V. (**2005**). Combining lectin microcolumns with high-resolution separation techniques for enrichment of glycoproteins and glycopeptides. *Anal Chem* **77**(13): 4081-90.

Madonna, A.J.; Basile, F.; Ferrer, I.*; et al.* (**2000**). On-probe sample pretreatment for the detection of proteins above 15 KDa from whole cell bacteria by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **14**(23): 2220-9.

Mamyrin, B.A.; Karataev, V.I.; Shmikk, D.A.*; et al.* (**1973**). The mass reflectron, a new non-magnetic time-of-flight mass spectrometer with high resolution. *Sov Phys JETP* **37**: 45-8.

Mann, M.; Meng, C.K.; Fenn, J.B. (**1989**). Interpreting mass spectra of multiply charged ions. *Anal Chem* **61**(15): 1702-8.

Mann, M.; Hojrup, P.; Roepstorff, P. (**1993**). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* **22**(6): 338-45.

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

Mann, M.; Wilm, M. (**1995**). Electrospray mass spectrometry for protein characterization. *Trends Biochem Sci* **20**(6): 219-24.

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

Mann, M.; Jensen, O.N. (**2003**). Proteomic analysis of post-translational modifications. *Nat Biotechnol* **21**(3): 255-61.

McDonald, L.; Robertson, D.H.; Hurst, J.L.*; et al.* (**2005**). Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat Methods* **2**(12): 955-7.

Medzihradszky, K.F.; Campbell, J.M.; Baldwin, M.A.*; et al.* (**2000**). The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal Chem* **72**(3): 552-8.

Meldrum, D. (**2000a**). Automation for genomics, part two: sequencers, microarrays, and future trends. *Genome Res* **10**(9): 1288-303.

Meldrum, D. (**2000b**). Automation for genomics, part one: preparation for sequencing. *Genome Res* **10**(8): 1081-92.

Meng, F.; Cargile, B.J.; Patrie, S.M.*; et al.* (**2002**). Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Anal Chem* **74**(13): 2923-9.

Morris, H.R.; Paxton, T.; Panico, M.*; et al.* (**1997**). A novel geometry mass spectrometer, the Q-TOF, for low-femtomole/attomole-range biopolymer sequencing. *J Protein Chem* **16**(5): 469-79.

Mortz, E.; O'Connor, P.B.; Roepstorff, P.*; et al.* (**1996**). Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases. *Proc Natl Acad Sci U S A* **93**(16): 8264-7.

Mukherji, M. (**2005**). Phosphoproteomics in analyzing signaling pathways. *Expert Rev Proteomics* **2**(1): 117-28.

Nagele, E.; Vollmer, M.; Horth, P. (**2003**). Two-dimensional nano-liquid chromatography-mass spectrometry system for applications in proteomics. *J Chromatogr A* **1009**(1-2): 197-205.

Nielsen, P.A.; Olsen, J.V.; Podtelejnikov, A.V.*; et al.* (**2005**). Proteomic mapping of brain plasma membrane proteins. *Mol Cell Proteomics* **4**(4): 402-8.

O'Farrell, P.H. (**1975**). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**(10): 4007-21.

Oehlers, L.P.; Perez, A.N.; Walter, R.B. (**2005**). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of 4-sulfophenyl isothiocyanate-derivatized peptides on AnchorChip sample supports using the sodium-tolerant matrix 2,4,6-trihydroxyacetophenone and diammonium citrate. *Rapid Commun Mass Spectrom* **19**(6): 752-8.

Oh-Ishi, M.; Satoh, M.; Maeda, T. (**2000**). Preparative two-dimensional gel electrophoresis with agarose gels in the first dimension for high molecular mass proteins. *Electrophoresis* **21**(9): 1653-69.

Oliver, B.; Leblanc, B. (**2003**). How many genes in a genome? *Genome Biol* **5**(1): 204.

Oliver, S. (**2000**). Guilt-by-association goes global. *Nature* **403**(6770): 601-3.

Olsen, J.V.; Ong, S.E.; Mann, M. (**2004**). Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics* **3**(6): 608-14.

Ong, S.E.; Blagoev, B.; Kratchmarova, I.*; et al.* (**2002**). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**(5): 376-86.

Paizs, B.; Suhai, S. (**2005**). Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* **24**(4): 508-48.

Pandey, A.; Mann, M. (**2000**). Proteomics to study genes and genomes. *Nature* **405**(6788): 837-46.

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

Pashkova, A.; Chen, H.S.; Rejtar, T.*; et al.* (**2005**). Coumarin tags for analysis of peptides by MALDI-TOF MS and MS/MS. 2. Alexa Fluor 350 tag for increased peptide and protein Identification by LC-MALDI-TOF/TOF MS. *Anal Chem* **77**(7): 2085-96.

Patterson, S.D.; Katta, V. (**1994**). Prompt fragmentation of disulfide-linked peptides during matrix-assisted laser desorption ionization mass spectrometry. *Anal Chem* **66**(21): 3727-32.

Pavy, N.; Rombauts, S.; Dehais, P.*; et al.* (**1999**). Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences. *Bioinformatics* **15**(11): 887-99.

Peng, J.; Gygi, S.P. (**2001**). Proteomics: the move to mixtures. *J Mass Spectrom* **36**(10): 1083-91.

Peng, J.; Elias, J.E.; Thoreen, C.C.*; et al.* (**2003**). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2**(1): 43-50.

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

Perou, C.M.; Sorlie, T.; Eisen, M.B.*; et al.* (**2000**). Molecular portraits of human breast tumours. *Nature* **406**(6797): 747-52.

Rabilloud, T. (**2000**). Detecting proteins separated by 2-D gel electrophoresis. *Anal Chem* **72**(1): 48A-55A.

Rabilloud, T. (**2002**). Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2**(1): 3-10.

Reed, J.L.; Famili, I.; Thiele, I.*; et al.* (**2006**). Towards multidimensional genome annotation. *Nat Rev Genet* **7**(2): 130-41.

Reinders, J.; Lewandrowski, U.; Moebius, J.*; et al.* (**2004**). Challenges in mass spectrometry-based proteomics. *Proteomics* **4**(12): 3686-703.

Reinders, J.; Sickmann, A. (**2005**). State-of-the-art in phosphoproteomics. *Proteomics* **5**(16): 4052-61.

Reinhold, V.N.; Reinhold, B.B.; Costello, C.E. (**1995**). Carbohydrate molecular weight profiling, sequence, linkage, and branching data: ES-MS and CID. *Anal Chem* **67**(11): 1772-84.

Resing, K.A.; Meyer-Arendt, K.; Mendoza, A.M.*; et al.* (**2004**). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem* **76**(13): 3556-68.

Righetti, P.G.; Campostrini, N.; Pascali, J.*; et al.* (**2004a**). Quantitative proteomics: a review of different methodologies. *Eur J Mass Spectrom* **10**(3): 335-48.

Righetti, P.G.; Castagna, A.; Antonucci, F.*; et al.* (**2004b**). Critical survey of quantitative proteomics in two-dimensional electrophoretic approaches. *J Chromatogr A* **1051**(1-2): 3-17.

Righetti, P.G.; Castagna, A.; Antonioli, P.*; et al.* (**2005a**). Prefractionation techniques in proteome analysis: the mining tools of the third millennium. *Electrophoresis* **26**(2): 297-319.

Righetti, P.G.; Castagna, A.; Herbert, B.*; et al.* (**2005b**). How to bring the "unseen" proteome to the limelight via electrophoretic pre-fractionation techniques. *Biosci Rep* **25**(1-2): 3-17.

Rochfort, S. (**2005**). Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. *J Nat Prod* **68**(12): 1813-20.

Roepstorff, P.; Fohlman, J. (**1984**). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **11**(11): 601.

Rose, J.K.; Bashir, S.; Giovannoni, J.J.*; et al.* (**2004**). Tackling the plant proteome: practical approaches, hurdles and experimental tools. *Plant J* **39**(5): 715-33.

Ross, P.L.; Huang, Y.N.; Marchese, J.N.*; et al.* (**2004**). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**(12): 1154-69.

Rouze, P.; Pavy, N.; Rombauts, S. (**1999**). Genome annotation: which tools do we have for it? *Curr Opin Plant Biol* **2**(2): 90-5.

Samyn, B.; Sergeant, K.; Castanheira, P.*; et al.* (**2005**). A new method for C-terminal sequence analysis in the proteomic era. *Nat Methods* **2**(3): 193-200.

Santoni, V.; Molloy, M.; Rabilloud, T. (**2000**). Membrane proteins and proteomics: un amour impossible? *Electrophoresis* **21**(6): 1054-70.

Schagger, H.; von Jagow, G. (**1991**). Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Anal Biochem* **199**(2): 223-31.

Schmidt, F.; Schmid, M.; Jungblut, P.R.*; et al.* (**2003**). Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *J Am Soc Mass Spectrom* **14**(9): 943-56.

Shaw, J.; Rowlinson, R.; Nickson, J.*; et al.* (**2003**). Evaluation of saturation labelling two-dimensional difference gel electrophoresis fluorescent dyes. *Proteomics* **3**(7): 1181-95.

Shevchenko, A.; Jensen, O.N.; Podtelejnikov, A.V.*; et al.* (**1996**). Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci U S A* **93**(25): 14440-5.

Smith, R.D.; Loo, J.A.; Edmonds, C.G.*; et al.* (**1990**). New developments in biochemical mass spectrometry: electrospray ionization. *Anal Chem* **62**(9): 882-99.

Sorlie, T.; Perou, C.M.; Tibshirani, R.*; et al.* (**2001**). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**(19): 10869-74.

Stein, L. (**2001**). Genome annotation: from sequence to biology. *Nat Rev Genet* **2**(7): 493-503.

Suckau, D.; Resemann, A. (**2003**). T3-sequencing: targeted characterization of the N- and C-termini of undigested proteins by mass spectrometry. *Anal Chem* **75**(21): 5817-24.

Sunyaev, S.; Liska, A.J.; Golod, A.*; et al.* (**2003**). MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* **75**(6): 1307-15.

Taher, L.; Rinner, O.; Garg, S.*; et al.* (**2004**). AGenDA: gene prediction by cross-species sequence comparison. *Nucleic Acids Res* **32**(Web Server issue): W305-8.

Tanaka, K.; Waki, H.; Ido, Y.*; et al.* (**1988**). Protein and polymer analyses up to m/z 100.000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **2**(2): 151-3.

Tao, W.A.; Aebersold, R. (**2003**). Advances in quantitative proteomics via stable isotope tagging and mass spectrometry. *Curr Opin Biotechnol* **14**(1): 110-8.

Thiede, B.; Hohenwarter, W.; Krah, A.*; et al.* (**2005**). Peptide mass fingerprinting. *Methods* **35**(3): 237-47.

Tjalsma, H.; Bolhuis, A.; Jongbloed, J.D.*; et al.* (**2000**). Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiol Mol Biol Rev* **64**(3): 515-47.

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

Vanrobaeys, F.; Devreese, B.; Lecocq, E.*; et al.* (**2003**). Proteomics of the dissimilatory iron-reducing bacterium Shewanella oneidensis MR-1, using a matrix-assisted laser desorption/ionization-tandem-time of flight mass spectrometer. *Proteomics* **3**(11): 2249-57.

Vanrobaeys, F.; Van Coster, R.; Dhondt, G.*; et al.* (**2005**). Profiling of Myelin Proteins by 2D-Gel Electrophoresis and Multidimensional Liquid Chromatography Coupled to MALDI TOF-TOF Mass Spectrometry. *J Proteome Res* **4**(6): 2283-93.

Verentchikov, A.N.; Ens, W.; Standing, K.G. (**1994**). Reflecting time-of-flight mass spectrometer with an electrospray ion source and orthogonal extraction. *Anal Chem* **66**(1): 126-33.

Verhaert, P.; Uttenweiler-Joseph, S.; de Vries, M.*; et al.* (**2001**). Matrix-assisted laser desorption/ionization quadrupole time-of-flight mass spectrometry: an elegant tool for peptidomics. *Proteomics* **1**(1): 118-31.

Vestal, M.L.; Juhasz, P.; Martin, S.A. (**1995**). Delayed extraction matrix-assisted laser desorption time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **9**(11): 1044-50.

Vestal, M.L.; Campbell, J.M. (**2005**). Tandem time-of-flight mass spectrometry. *Methods Enzymol* **402**: 79-108.

Vollmer, M.; Hörth, P.; Nägele, E. (**2003**). Optimization of two-dimensional off-line LC/MS separations to improve resolution of complex proteomic samples. *Anal Chem* **76**(17): 5180-5.

Walhout, A.J.; Vidal, M. (**2001**). Protein interaction maps for model organisms. *Nat Rev Mol Cell Biol* **2**(1): 55-62.

Wang, H.; Hanash, S. (**2003**). Multi-dimensional liquid phase based separations in proteomics. *J Chromatogr B* **787**(1): 11-8.

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**(3): 242-7.

Wiley, W.C.; McLaren, I.H. (**1955**). Time-of-Flight mass spectrometer with improved resolution. *Rev Sci Instru* **26**(12): 1150-7.

Wilkins, M.R.; Sanchez, J.C.; Gooley, A.A.*; et al.* (**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.

Williamson, B.; Marchese, J.; Juhasz, P.*; et al.* (**2002**). Diving deeper into the proteome: next generation ICAT™ reagents coupled with MDLC and high-throughput MS. *Proceedings of the 50th Conference on Mass Spectrometry and Allied Topics, Orlando, Fl*.

Wilm, M.; Shevchenko, A.; Houthaeve, T.*; et al.* (**1996**). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**(6564): 466-9.

Wolfe, C.J.; Kohane, I.S.; Butte, A.J. (**2005**). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics* **6**: 227.

Wu, C.C.; MacCoss, M.J.; Howell, K.E.*; et al.* (**2003**). A method for the comprehensive proteomic analysis of membrane proteins. *Nat Biotechnol* **21**(5): 532-8.

Wu, W.W.; Wang, G.; Baek, S.J.*; et al.* (**2006**). Comparative study of three proteomic quantitative methods, DIGE, cICAT, and iTRAQ, using 2D gel- or LC-MALDI TOF/TOF. *J Proteome Res* (accepted for publication).

Yamada, N.; Suzuki, E.; Hirayama, K. (**2006**). Effective novel dissociation methods for intact protein: Heat-assisted nozzle-skimmer collisionally induced dissociation and infrared multiphoton dissociation using a Fourier transform ion cyclotron resonance mass spectrometer equipped with a micrometal electrospray ionization emitter. *Anal Biochem* **348**(1): 139-47.

Yang, X.; Wu, H.; Kobayashi, T.*; et al.* (**2004a**). Enhanced ionization of phosphorylated peptides during MALDI TOF mass spectrometry. *Anal Chem* **76**(5): 1532-36.

Yang, Z.; Hancock, W.S. (**2004b**). Approach to the comprehensive analysis of glycoproteins isolated from human serum using a multi-lectin affinity column. *J Chromatogr B* **1053**(1-2): 79-88.

Yao, X.; Freas, A.; Ramirez, J.*; et al.* (**2001**). Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**(13): 2836-42.

Yates, J.R., 3rd; Speicher, S.; Griffin, P.R.*; et al.* (**1993**). Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem* **214**(2): 397-408.

Zhang, H.; Yan, W.; Aebersold, R. (**2004**). Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr Opin Chem Biol* **8**(1): 66-75.

Zhong, H.; Zhang, Y.; Wen, Z.*; et al.* (**2004**). Protein sequencing by mass analysis of polypeptide ladders after controlled protein hydrolysis. *Nat Biotechnol* **22**(10): 1291-6.

Zhou, G.; Li, H.; DeCamp, D.*; et al.* (**2002**). 2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers. *Mol Cell Proteomics* **1**(2): 117-24.

Zubarev, R.; Hakansson, P.; Sundqvist, B. (**1996**). Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Anal Chem* **68**(22): 4060-63.

PART 2

C-TERMINAL SEQUENCE ANALYSIS IN THE PROTEOME ERA

2.1. Introduction

2.1.1. Why C-terminal sequence analysis?

Despite the number of methods that have been developed for the analysis of proteins, questions concerning the C-terminus of proteins are frequently left unanswered. Nevertheless, the C-terminus of proteins is often an important determinant of their function. In general, the C-terminus is situated on the outside of proteins in contact with the solvent. Therefore, the C-terminal amino acid, together with a number of proceeding residues, is ideally situated to serve as a recognition motif capable of conferring a variety of essential biochemical reactions. Some of the known functions that are determined by the C-terminal sequence include protein trafficking, subcellular anchoring of proteins, targeted protein degradation, and the static and dynamic formation of macromolecular complexes (Chung *et al*, 2002; Chung *et al*, 2003). Performing these different functions requires the formation of stable interactions; C-terminal sequence motifs (Gatto *et al*, 2003) and protein domains that specifically interact with them are known. One example is the clustering of ion membrane channels through interaction of their C-terminal sequence (X-Ser/Thr-X-Val-COOH) with modular PDZ domains, quoted as the 'figurative glue' that keeps protein complexes together (Doyle *et al*, 1996; Fanning *et al*, 1999). Since C-terminal sequence motifs can occur in different proteins within a given species, they also code for specific biological activities among structural and functional distinct proteins. The presence of a K/HDEL C-terminal sequence motif is a recognition marker that destines proteins for retrieval or retention in the endoplasmic reticulum, irrespective of the function they have to perform in that cell compartment (Jurgens, 2004).

Proteolytic processing plays an important role in determining the function of many proteins. The most familiar form of proteolytic processing is the cleavage of signal peptides after transmembrane trafficking of proteins (Chou, 2001; Krimmer *et al*, 2001). Apart from N-terminal processing, C-terminal proteolytic processing is of importance in a wide variety of biological events. The generation of peptide hormones from inactive prohormones is one of the complex pathways involving both N- and C-terminal processing (van Strien *et al*, 1996). Depending on external stimuli, proopiomelanocortin (POMC) is processed at dibasic sites by prohormone convertases (PCI & PCII) (Tanaka, 2003). After trimming by carboxypeptidases, the different peptides resulting from these cleavages are responsible for the regulation of diverse physiological responses such as pain suppression and the regulation of food uptake (Raffin-Sanson *et al*, 2003). For most of these peptide hormones, the processing sites, their physiological effects, and the regulation of their excretion is known. However, the function of some is yet unknown. Malignant proteolytic processing, apart from the normal 'healthy' one, can result in the occurrence of diseases. One of the most supported hypotheses for the development of plaques in Alzheimer's disease, the amyloid hypothesis (Hardy *et al*, 2002), is that aberrant processing of the amyloid precursor protein (APP) finally leads to the formation of plaques and neuronal cell death. The activity of β-secretases and subsequent heterogeneous cleavage by a multi-enzyme complex (γ-secretase) results in a mixture of peptides of which amyloid β40 and β42 peptides are the most abundant components. After seeding, formation of small plaques, accumulation of β-amyloid peptides triggers a cascade that ultimately leads to cell death (Jarrett *et al*, 1993). Despite the fact that the infectious character of prion-associated pathologies isolates them from age-related neurodegenerative syndromes, the link between prion diseases and Alzheimer's is apparent. The cellular prion protein is processed by proteases from the disintegrin family as is the amyloid precursor protein. Furthermore, similar signal transduction cascades ultimately result in cell death (Checler *et al*, 2002).

Wilkins *et al* demonstrated that short C-terminal sequences (sequence tag) can be used in database searches, resulting in a higher specificity than the use of a N-terminal tag of the same length (Wilkins *et al*, 1998). An efficient method for C-terminal sequence analysis could also have an impact on biotechnology. The knowledge of both termini of a protein allows a more efficient choice of primers for the cloning of genes from species whose genome is not known. Finally, C-terminal sequence analysis can aid in quality control of protein products after cloning and over-expression of genes: it ensures that the gene product is formed completely (Ciurli *et al*, 2002; Zambelli *et al*, 2005).

2.1.2. Chemical C-terminal sequence analysis

Although several methods for chemical C-terminal sequence analysis have been developed (Ward, 1986; Inglis, 1991), none of these methods reached the level of efficiency that is routinely attained using N-terminal Edman degradation. The nucleophilic primary-amino group of proteins readily reacts with PITC under mild conditions, resulting in high repetitive yields of 95-99%. For C-terminal sequence analysis there is no group with similar reactivity available, thus necessitating the development of chemical methods that convert the poor nucleophilic carboxyl group to a more reactive functionality. Only the use of the so called thiocyanate degradation has some success.

The thiocyanate degradation consists of three steps; activation, derivatization and cleavage. During the activation, the C-terminal carboxyl group is converted to a reactive electrophilic group such as a mixed anhydride or an oxazolone. This group is subsequently derivatized with a thiocyanate reagent to a thiohydantoin (Figure 2.1). Finally, the thiohydantoin is specifically cleaved off and analyzed by reversed phase HPLC. The conversion of amino acids to thiohydantoins was described for the first time in 1911 (Johnson *et al*, 1911). The sequential application of this derivatization on peptides was made possible by adding a cleavage step (Schlack *et al*, 1926). From this point on, different reagents for activation and cleavage have been used. Currently, two different approaches have been automated by Applied Biosystems and Hewlett-Packard (Henry, 1998).

The automated C-terminal sequencer from Hewlett-Packard uses DPP-ITC (diphenyl phosphoroisothiocyanatidate) to derivatize the C-terminal amino acid. The addition of pyridine to the derivatization reaction allows the complete conversion to a thiohydantoin in less then 60 minutes (Bailey *et al*, 1992). After reaction with DPP-ITC pyridine is delivered in the gas phase, resulting in the rearrangement of the pentavalent acylphosphoryl isothiocyanate to the acyl isothiocyanate, which spontaneously rearranges to a thiohydantoin. The thiohydantoin is subsequently released through reaction with sodium trimethylsilanolate (Bailey *et al*, 1994). It is claimed that, using this chemistry, all natural occurring amino acids can be converted to a thiohydantoin. However, in the case of proline this would require the formation of a quaternary amide nitrogen. Reaction of the proline thiohydantoin with the cleavage reagent results in the regeneration of the C-terminal proline, thereby effectively blocking the peptide for further sequence analysis. A more stable derivative, protonated proline thiohydantoin, can be formed through reaction with a strong acid and hydrolysis with water vapor. These two steps, reaction with acid and hydrolysis with water vapors, have no influence on the derivatization of the other amino acids and can be inserted in the standard program. Because of its volatility, TFA is the preferred acid for the cleavage; it can even be delivered in the gas phase. When applied on proteins having a C-terminal proline, sequence analysis of this amino acid was possible albeit with relatively low yields (0,5 to 5%) (Bailey, 1995; Bailey *et al*, 1995). In general, the initial yield varies between 10 and 50%. Together

with the moderate repetitive yield, this allows the determination of about 5 C-terminal amino acids starting from nanomolar amounts of protein (Miller *et al*, 1995).

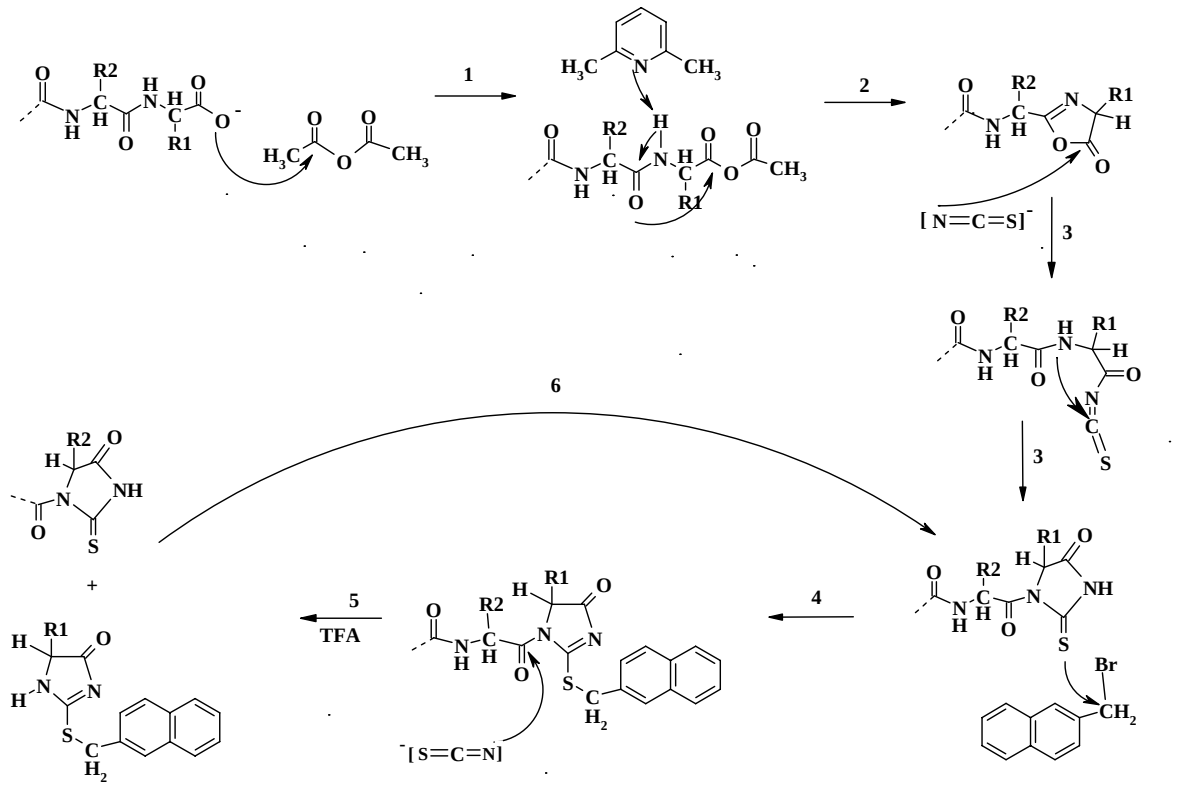


Figure 2.1. Chemical C-terminal sequence analysis using the alkylation chemistry. In steps 1 to 3, the C-terminal amino acid is activated and derivatized to a thiohydantoin by reaction with respectively acetic acid anhydride and a thiocyanate anion from tetrabutylammonium thiocyanate. Alkylation of the sulfur in the thiohydantoin ring with bromomethyl naphthalene converts the amino acid derivative to a better leaving group. Cleavage using a thiocyanate anion results in the formation of an ATH-derivative. The penultimate amino acid is converted to a thiohydantoin; therefore, only step 4 and 5 must be repeated to determine the following residues.

In 1992 a new method was described, in which the sulfur atom in the thiohydantoin ring is alkylated with bromomethyl naphthalene, converting the thiohydantoin to a better leaving group (Boyd *et al*, 1992). The alkylated thiohydantoin (ATH) can subsequently be cleaved using an isothiocyanate reagent that simultaneously converts the penultimate amino acid to a thiohydantoin. The concomitant performance of cleavage and derivatization of the next amino acid implies that the activation cycles must be performed only once (Figure 2.1). The thiohydantoin shows structural resemblance to the PTH-derivative of the Edman degradation, but there is a difference in reactivity and chemical stability. Deprotonation of the N_3 results in the formation of a resonance stabilized anion having a relatively low pKa-value (about 7). Therefore, the peptidyl-thiohydantoin can be alkylated in basic environments. Because of their high reactivity and the fact that their aromatic system improves the UV-absorbing properties of the ATHs, benzylbromide derivatives are preferred as alkylating reagents. Cleavage of the ATH is acid catalyzed and can also be performed with TFA alone, but the resulting free carboxyl group then requires that the entire cycle (activation, derivatization and cleavage) is repeated for every amino acid.

The inability to sequence through proline is one of the main problems in the use of the alkylation chemistry for C-terminal sequence analysis. It was shown that the derivatization of N-acetylproline to proline thiohydantoin is possible with a yield of 100% (Hardeman *et al*,

1998). Nevertheless, currently C-terminal sequence analysis through proline with this method remains to be proven.

2.1.3. Mass spectrometric methods for C-terminal sequence analysis

Whereas methods for N-terminal ladder sequence analysis are mainly based on chemical ladder-generating procedures (Chait *et al*, 1993), proteolytic digestion is the preferred method for C-terminal ladder sequence analysis (Bergman, 2000). In these methods peptides are incubated with one or a mixture of carboxypeptidases, exopeptidases that progressively remove amino acids from the C-terminus of polypeptides. A wide variety of exopeptidases can be used for the incubation but, because of their broad amino acid specificity, the most popular are the carboxypeptidases from bakers yeast (CPY) and from the fungus *Penicillum janthinellum* (CPP). Notwithstanding the broad specificity of these enzymes, some sequences are refractory to digestion (Breddam, 1986). Both CPP and CPY are unable to truncate very short peptides (2 to 3 amino acids). Thereby suggesting that the specificity of these exoproteases is not only determined by the C-terminal amino acid, but that amino acids adjacent to the C-terminus influence the C-terminal digestion. However, no exhaustive studies to determine the specificity of these carboxypeptidases have been performed. Since not all amino acids are removed at the same speed, multiple analyses must be performed to determine longer stretches of sequence. This can be achieved in a time-dependent or concentration-dependent experiment (Patterson *et al*, 1995). In time-dependent analysis, aliquots are taken from the reaction mixture at different time points and analyzed by mass spectrometry. In concentration-dependent digestions, samples are analyzed after parallel microdigestions on the probe using carboxypeptidase solutions of different concentrations. At a certain time point, matrix solution is added, thereby stopping the digestion, and the sample is prepared for analysis using MALDI-MS. Although mass spectrometry is the method of choice for the analysis of C-terminal sequence ladders, the removed C-terminal amino acids were initially analyzed by amino acid analysis. Different types of mass spectrometers have been used for the analysis of time dependent carboxypeptidase digestions; PD-MS (Klarskov *et al*, 1989), ESI-MS (Rosnack *et al*, 1992), MALDI-MS (Thiede *et al*, 1995; Bonetto *et al*, 1997) and SELDI-MS (Cool *et al*, 2004). Recently, the digestion with carboxypeptidases has also been performed using a continuous flow micro total analysis system (μ-TAS). In this proof-of principle study promising results showed that the automated system allows fast and sensitive analyses (Brivio *et al*, 2002).

An oxazolinone is postulated to be an intermediate in chemical methods for C-terminal ladder sequence analysis. Hydrolysis of peptides with vapors of fluorinated organic acids or their corresponding anhydrides results in the specific degradation of C-terminal amino acids (Tsugita *et al*, 1992). Because incubation of peptides in acids results in the cleavage of acid-labile bonds (i.e. at the C-terminal side of aspartic acid and the N-terminal side of serine), this method simultaneously results in the determination of internal sequence (Tsugita *et al*, 1998). Although the truncation with perfluoric acid anhydride is not easy to perform, it suppresses the cleavage of internal peptide bonds and, as such, allows more specific C-terminal sequence analysis (Takamoto *et al*, 1995). Reactive groups can be protected from side reactions through acetylation prior to the truncation reaction. Recently this approach was used to determine the C-terminal sequence of intact proteins. After acetylation, truncation and hydration the protein is digested with trypsin. This allowed the direct determination of short C-terminal sequences from myoglobin and trypsin inhibitor (Miyazaki *et al*, 2004).

2.1.4. Rationale for the development of a new method

Conceptually, the simplest way to determine the length, and thus the C-terminus of a protein is the analysis of its mass. Naturally, the sequence of the protein must be known and no unexpected mass shifts may be present (posttranslational modifications, N-terminal truncation, adduct formation, …). Analysis of intact proteins with FTMS results in a high resolution and mass accuracy and thus allows the efficient determination of the mass of intact proteins. Fragmentation of intact proteins followed by analysis of the fragmentation spectra has been applied for C-terminal sequence determinations and provides more flexibility than the determination of the mass alone. Fragmentation of intact proteins has also been described using other types of mass spectrometers: Applied Biosystems 4700 TOF/TOF (Lin *et al*, 2003), Bruker MALDI LIFT-TOF/TOF (Suckau *et al*, 2003) and Micromass Q-TOF (Rai *et al*, 2002). Although this approach is promising (Du *et al*, 2004; Patrie *et al*, 2005), problems in the analysis of complex samples are mainly situated in the front-end sample handling steps.

In chemical approaches, ladder sequence analysis and automated chemical sequence analysis, abundant side reactions take place. Incubation of peptides with perfluoric acid or the corresponding anhydride results in reactions on side-chain amino groups. Complete acetylation of the protein prior to the truncation reaction and removal of acetylations by hydration with vapors of 10% dimethylamino-ethanol (DMAE) lowers the amount and the extent of these side reactions. However, ion series with mass shifts characteristic for the respective side reactions are still observed in the mass spectra (Miyazaki *et al*, 2004). Using this type of reaction, no truncation of the peptide chain is observed when proline is the C-terminal amino acid (unpublished results). Furthermore, a sensitivity allowing the analysis of samples containing less than 100 pmol purified peptide/protein has never been reported. Prior and during the first cycle of automated chemical sequence analysis, specific modifications are performed to avoid the occurrence of reactions on the side chains of Cys, Lys, Ser, Thr, Glu and Asp. Samyn *et al* proved that the use of an optimized chemistry results in a sensitivity comparable to Edman degradation (Samyn *et al*, 2000). However, the low repetitive yield, about 20%, precludes the determination of long stretches of C-terminal sequence. The application of an improved activation chemistry, capable of derivatizing all natural occurring amino acids (Hardeman *et al*, 1998), has not been realized yet but is promising for C-terminal sequence analysis through proline.

In vitro, carboxypeptidases are only active on peptides, or at least partially denatured proteins. In a study on the deletion of the C-terminal region of the endopolygalacturonase from the fungus *Stereum purpureum,* an incubation time of 60 minutes was needed to determine the C-terminal amino acid (Shimizu *et al*, 2000). In earlier experiments, we demonstrated that up to 10 amino acids could be determined from the C-terminus of apocytochrome c after less than 20 minutes incubation with a cocktail of carboxypeptidases (Figure 2.2). When the same protocol was applied on native cytochrome c or on other intact proteins, we were unable to determine the C-terminal sequence (unpublished results).

Carboxypeptidase incubation of peptide mixtures, obtained after endoprotease digestion of proteins, results in overlapping sequence ladders from the different peptides. One method to overcome this limitation is the specific isolation of the C-terminal peptide from peptide samples. Anhydrotrypsin, a catalytically inert trypsin derivative wherein the catalytic serine is converted to dehydroalanine, allows the isolation of C-terminal peptides from samples after digestion of the proteins with trypsin and endoprotease LysC (Ishii *et al*, 1983). Using the strong affinity of anhydrotrypsin for peptides containing a C-terminal arginine or

lysine, these peptides can be selectively isolated. C-terminal peptides, not ending in Lys or Arg, are recovered from the flow-through fraction while internal peptides are immobilized (Kumazaki *et al*, 1986). In 2000, a more sensitive method was described in which anhydrotrypsin was coupled to agarose beads (Sechi *et al*, 2000). After isolation of the C-terminal peptide, it can be incubated with carboxypeptidases or be fragmented using tandem MS. Nevertheless, anhydrotrypsin was never used at a sensitivity compatible with 2D-gels or gel-free proteomic protocols. The selection of C-terminal peptides appears to be less than 100% specific, and can not be applied on proteins with a basic C-terminal amino acid. Therefore, the bias for basic residues, for Lys to a higher degree than for Arg, as C-terminal amino acid of proteins limits the applicability of this approach (Pal *et al*, 2000). A recent study indicated that between 23% (*Methanococcus jannaschii*) and 13% (*Homo sapiens*) of the predicted proteins of an organism have a C-terminal Lys or Arg (Gatto *et al*, 2003).

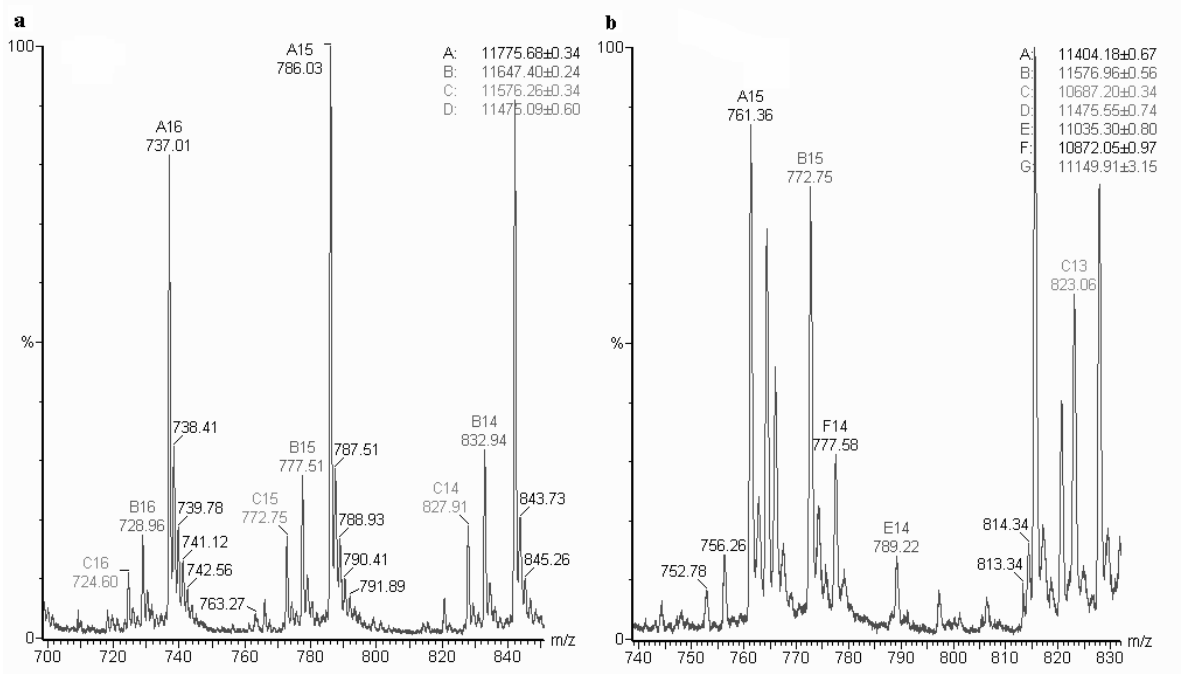


Figure 2.2. Ladder sequence analysis of pigeon apocytochrome c after CPX incubation. At certain time points, aliquots were taken, acidified to stop the digestion and analyzed using a ESI Q-TOF; a) after 1 minute and b) after 20 minutes of incubation at room temperature. Using these and other masses of the truncated protein at time points between 1 and 20 minutes the C-terminal sequence –[I/L]AY[I/L][K/Q][K/Q]ATA[K/Q] was deduced.

Proteolytic ^{18}O-labeling can be used to distinguish C-terminal peptides from the bulk of peptides after proteolysis of a protein (Kosaka *et al*, 2000). During hydrolysis of peptide bonds, oxygen from the solvent is incorporated in the carboxyl group of the resulting peptides. In proteolytic labeling approaches, complete equilibration of oxygen ($^{16}O/^{18}O$) between peptide and solvent is possible using some endoproteases (Schnolzer *et al*, 1996). A 2 Da mass shift is observed for all but the C-terminal peptide. Kosaka *et al* used a high resolution FTMS for this purpose but similar results can be obtained using other mass spectrometers. However, because a less than 50% sequence coverage is routinely attained in peptide mass fingerprints, the C-terminal peptide is not always observed. Furthermore, because of the repeating binding-hydrolysis cycles of trypsin (Schnolzer *et al*, 1996) this method does not allow the recognition of C-terminal peptides if the protein has a C-terminal Lys or Arg.

2.1.5. Research strategy

In the new method that we developed (Part 2.2), we use a combination of chemical protein cleavage and carboxypeptidase digestion. Incubation of proteins with cyanogen bromide (CNBr) results in the cleavage C-terminal to methionine. During this cleavage reaction, methionine is converted to a homoserine-derivative which is in equilibrium with its lactone (Figure 2.3). We demonstrated that peptides ending on homoserine lactone are resistant to carboxypeptidase digestion. Therefore, during incubation of the peptide mixture, only the original C-terminal peptide will be degraded (ladder formation) whereas no ladder is formed from internal fragments.

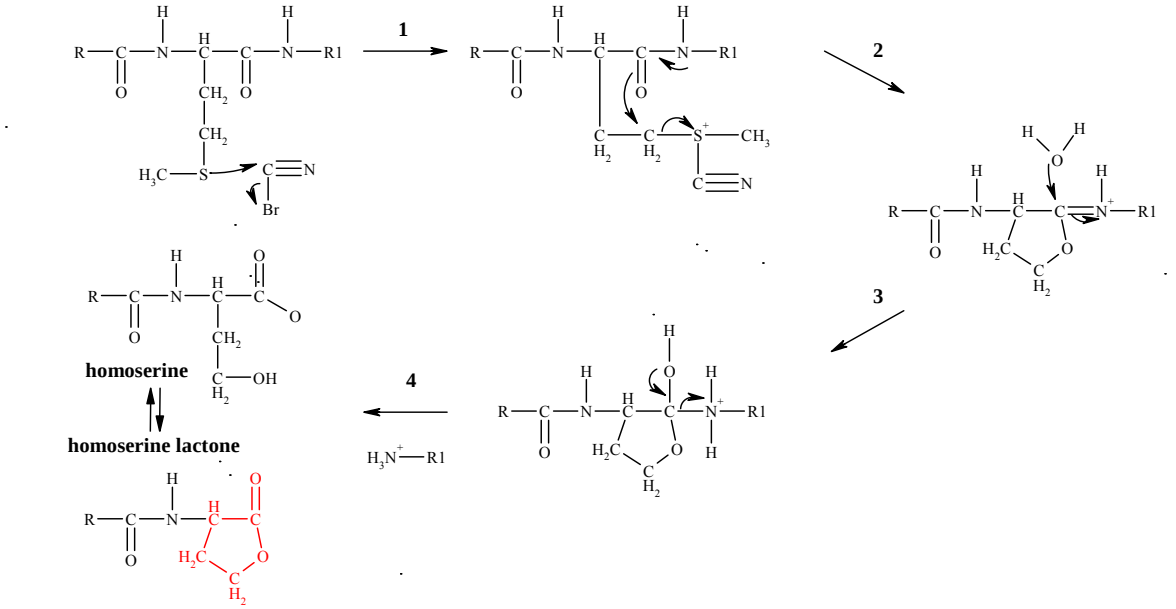


Figure 2.3. CNBr cleavage reaction. Incubation of proteins with CNBr results in cleavage of peptide bonds C-terminal to methionine. During the cleavage reaction methionine is converted to a homoserine lactone in equilibrium with its open form. The bromide-stabilized *S*-cyanide methionine derivative (Step 1) is susceptible to nucleophilic attack of the peptide bond carbonyl (Step 2). This acid catalyzed rearrangement results in the formation of a 5-membered cyclic structure and the leaving of methylisothiocyanate. The cyclic structure is resolved by nucleophilic attack from water, resulting in peptide bond cleavage and the formation of homoserine lactone (Step 3 & 4). Nucleophilic attack by water is impaired by competition with hydroxyl groups from threonine or serine when these amino acids are preceded by methionine.

The use of CNBr as cleavage agent for polypeptides was first proposed in 1962 (Gross *et al*, 1962). Because leaving of methylisothiocyanate (Figure 2.3, step 2) is an acid catalyzed reaction, CNBr-cleavage is always performed in acidic environments. All reagents used in this protocol (0.1% HCl and CNBr) are volatile, a property that is maintained using other reaction mixtures for CNBr cleavage, 70% formic acid (Compagnini *et al*, 2001) and 70% TFA. Incubations in 70% TFA (Shively *et al*, 1982) are beneficial, as formylation of reactive side chains during the cleavage reaction is avoided (Goodlett *et al*, 1990). The cleavage yield using 70% TFA is nearly 100%, except if the residue after methionine is serine or threonine (Morrison *et al*, 1990). The side chain hydroxyl group of these amino acids can interfere with the cleavage reaction by acting as nucleophile and competes with water in the cleavage reaction (Figure 2.3, step 3). To avoid this, the reaction can be performed in solutions containing a higher percentage of water, consequently a lower concentration of acid (Kaiser *et al*, 1999). For CNBr-cleavage, a strict requirement is that methionine residues are not oxidized; oxidation of methionine is a spontaneous reaction that totally impairs the reaction. Despite the harsh reaction conditions, in situ CNBr cleavage of proteins on PVDF- (Scott *et*

al, 1988) and nitrocellulose-membranes (Dukan *et al*, 1998) or in-gel (Loo *et al*, 1996; Cordoba *et al*, 1997) has been demonstrated.

The equilibrium between homoserine lactone and its open form is pH sensitive, and the two forms can be quantitatively converted in each other (Ambler, 1965). In acidic environments, the lactone form is preferred over the open form. Incubation of peptides with a C-terminal homoserine in TFA completely converts these to the lactone form in less then 60 minutes. The buffer we use for carboxypeptidase digestion has an acidic pH, therefore the equilibrium is almost completely shifted to the lactone. The specific reactivity of the homoserine lactone has been used in different approaches to recognize the C-terminal peptide of a protein. Comparison of the mass spectra of a CNBr peptide mixture before and after reaction with acidic methanol revealed a mass shift of 32 Da for all but the C-terminal peptide (Murphy *et al*, 1995). The specific reactivity of tris(hydroxymethyl)aminomethane (Tris) with homoserine lactones can be used in a similar way (Compagnini *et al*, 2001).

Carboxypeptidase P en Y are both serine exoproteases with a broad substrate specificity. These carboxypeptidases are remarkable among the serine protease family as they have a very low pH-optimum, between 2.5 and 5.7 for CPP and between 5.5 and 7.5 for CPY. In studies in which the residues surrounding the active site were mutated, it was shown that this low pH optimum depends on the ionization of the essential histidine in the serine protease catalytical triad (Bech *et al*, 1989). A model has been suggested to explain the low pKa of the essential histidine (Figure 2.4) (Stennicke *et al*, 1996). The C-terminal peptide bond of the substrate is polarized through the formation of hydrogen bridges between the substrate and the enzyme active site. The partial double bond character of the peptide bond is lost and the amino group is able to accept a hydrogen atom from the essential histidine, while the oxyanion is stabilized. The deprotonated histidine can now act as a nucleophile and the remainder of the reaction mechanism is the same as for other serine proteases. The pH optimum for the hydrolysis of C-terminal peptide esters, a bond that can not be polarized via this mechanism, is higher than for the hydrolysis of peptide bonds.

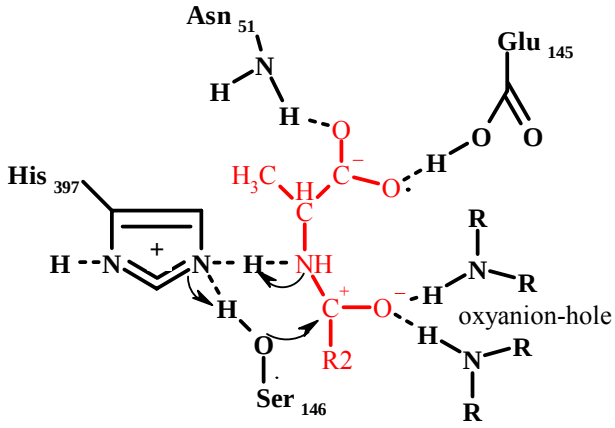
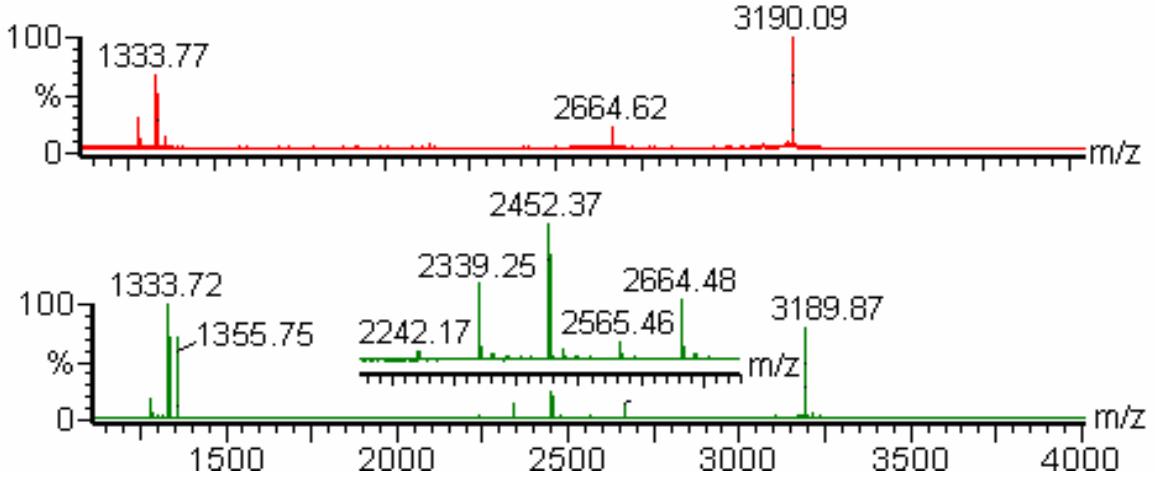


Figure 2.4. Activation of the essential histidine in CPP and CPY. The C-terminal peptide bond of the substrate (red) is polarized and hydrogen transfer from the essential histidine (His_{397}) to the substrate allows this histidine to activate the active site serine (Ser_{146}) as in other serine proteases.

The broad substrate specificity of carboxypeptidase Y and P can be explained from their 3D-structure. Flexibility is observed when the 3D-structures of CPY with and without substrate are compared (Endrizzi *et al*, 1994). In the study presented here, each of the 20 natural occurring amino acids were cleaved off. Furthermore, the cleavage of a chemically modified cysteine, carboxyamidomethyl-cysteine, at rates comparable to other amino acids illustrates the broad substrate specificity. Nevertheless, preferences for both the C-terminal and the penultimate amino acid exist. Hydrophobic residues are preferred at the penultimate position; Phe>Leu>Ala>Ile>>>Lys for both carboxypeptidases (Breddam, 1986). CPY has a lower cleavage rate for Gly and Asp as C-terminal amino acid while incubation with CPP alone results in a slower cleavage rate when Ser or Gly are C-terminal. Incubation of peptides

with a combination of CPP and CPY overcomes the specificity of the individual carboxypeptidases. Because of this more homogenous cleavage rate, longer stretches of sequence can be obtained when such a combination is used (Thiede *et al*, 1995). Why the homoserine lactone is not cleaved by these carboxypeptidases is not yet explained. However, it is this lack of cleavage that allows the determination of the C-terminal sequence of a protein by incubation of the mixture of CNBr-peptides with carboxypeptidases. Only the original C-terminal peptide is degraded and sequence ladders in the mass spectrum after incubation with carboxypeptidases are uniquely from the C-terminal peptide (Figure 2.5).



Peptide[a]	Mass (Da)[a]	Exp mass (Da)	Sequence[a]
1-93	105025.8	/	RELEELNVPGEIVESLSSSEESITRINKKIEKFQSEEQQQTEDELQDKI HPFAQTQSLVYPFPGPIPNSLPQNIPPLTQTPVVVPPFLQPEVhsl
94-102	900.1	/	GVSKVKEAhsl
103-109	792.0	/	APKHKEhsl
110-144	4020.6	/	PFPKYPVEPFTESQSLTLTDVENLHLPLPLLQSWhsl
145-156	1333.6	1333.7	HQPHQPLPPTVhsl
157-185	3189.6	3189.9	FPPQSVLSLSQSKVLPVPQKAVPYPQRDhsl
186-209	2664.2	2664.5	PIQAFLLYQEPVLGPVRGPFPIIV

Figure 2.5. C-terminal sequence analysis of bovine β-casein. Upper mass spectrum; after cleavage of β-casein with CNBr, 3 peptides are observed; the remainder of the sequence is found in 4 peptides that are outside the optimal mass range of the used mass spectrometer (M@LDI, Micromass). After 10 minutes of incubation with carboxypeptidases, a sequence ladder is formed from the C-terminal peptide (2664.53 Da) from which the C-terminal sequence, PIIV, can be determined (inset). No C-terminal truncation is observed from the other peptides. The table lists the masses of the peptides after CNBr-digestion of bovine β-casein. All but the original C-terminal peptide have homoserine lactone (hsl) as C-terminal amino acid. The absence in the spectra of peaks at +18 Da indicates that the homoserine-homoserine lactone equilibrium is completely shifted towards the lactone form. [a] based on NCBI database entry 115660, with the signal peptide removed.

2.2. C-terminal sequence analysis in the proteome era

A novel method for C-terminal sequence analysis in the proteomic era

Bart Samyn[1], Kjell Sergeant[1], Pedro Castanheira[2], Carlos Faro[2] and Jozef Van Beeumen[1]

[1] University Ghent
Department Biochemistry, Physiology and Microbiology
Laboratory of Protein Biochemistry and Protein Engineering
K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

[2] Departamento de Bioquimica and Centro de Neurociências e Biologia Celular
Universidade de Coimbra
Apt. 3126, 3000 Coimbra, Portugal

Published in *Nature Methods* (**2005**), 2(*3*), 193-200
J Biol Chem (**2005**), 280(*13*), 13047-54

Introduction

Mass spectrometry (MS) has emerged as a key platform technology in proteomics. In most approaches, a few peptides are identified which is sufficient to identify the protein but fails to characterize the exact state of the gene product, including post-translational modifications (PTMs) (Rappsilber *et al*, 2002). The variety of methodologies used for PTMs analysis has recently been reviewed (Mann *et al*, 2003). Most of the strategies are targeted to specific types of modifications such as phosphorylation and glycosylation (Dove, 2001; Zhou *et al*, 2001; Knight *et al*, 2003). However, considerably less attention has been given to study other kinds of PTMs. More specifically, the study of N- and C-terminal proteolytic processing is hampered by the unavailability of suitable methods (Mann *et al*, 2003).

Currently, the only methods that allow direct confirmation of the termini of proteins are chemical degradation techniques. Amino-terminal protein sequencing by Edman degradation is still the method of choice to determine N-terminal proteolytic processing. A chemical method for carboxy-terminal sequence analysis should provide a complementary approach. Such methods have been developed and automated (Bailey *et al*, 1992; Boyd *et al*, 1992), but there remain a number of unsolved problems, such as their modest sensitivity (20-100 pmol) and their low repetitive yields (Samyn *et al*, 2000). Furthermore, several amino acids (AA) require chemical modification and it remains difficult to sequence through proline (Bailey, 1995; Hardeman *et al*, 1998).

Characterization of the C terminus by MS fragmentation of intact proteins, the "top-down" approach, has been demonstrated on a variety of instruments (Aebersold *et al*, 2003). However, due to difficulties in dealing with whole proteins, these approaches are not in common use and their full exploitation awaits technical improvements (Meng *et al*, 2001; Meng *et al*, 2002; Suckau *et al*, 2003). Alternatively, methods have been described that allow the isolation of the C-terminal peptide (Kosaka *et al*, 2000; Sechi *et al*, 2000). However, they have rarely been applied, most likely due to problems associated with the recovery of the peptide and the need for larger sample amounts (Kosaka *et al*, 2000; Zhou *et al*, 2004).

In the so-called "ladder sequencing" techniques (Chait *et al*, 1993), a sequence-defining set of peptide fragments, each differing from the next by a single residue, is generated and analyzed by MS. Whereas chemical ladder generating procedures have mainly been developed for N-terminal approaches, proteolytic digestion is the preferred method for C-terminal analysis of peptides.

Here, we report a new approach directed at the selective characterization of the C-terminal sequence. The MS-based, enzymatic, ladder sequencing approach was applied on the unseparated peptide mixture, generated by cleavage of the protein with cyanogen bromide (CNBr). During incubation with carboxypeptidases (CP) only the original C-terminal fragment is accessible to enzymatic degradation and forms a ladder. Ladder read-out was performed using a MALDI-TOF/TOF instrument. The rate at which AA are cleaved by carboxypeptidases depended to a great extent on the peptide sequence. In experiments where only a few C-terminal residues are removed, indicating that this peptide represents the C-terminal fragment, the peptide was subjected to MALDI MS/MS fragmentation. The different steps in the protocol were optimized using a panel of standard proteins and further improved for analysis of gel-separated proteins. To illustrate the applicability of this approach at a proteomic scale, we characterized a number of C-terminal sequences from two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) separated proteins of the bacterium

Shewanella oneidensis. Finally, we used the method to investigate the proteolytic processing of a plant procardosin A.

Methods

Test protein mixture

A set of standard proteins was prepared containing the following proteins (Sigma): pigeon cytochrome c (PCC)(96% pure), horse heart myoglobin (MYO)(90% pure), bovine β-casein (90% pure); all of these proteins have no disulfide bridges and do not require reduction and S-alkylation prior to digestion. β-Lactoglobulin (BLG)(100%) was from Applied Biosystems. Bovine ribonuclease B (pancreatic, 180 Kunitz U/mg protein), bovine serum albumin (BSA)(98% pure) and yeast alcohol dehydrogenase (YADH-1)(90% pure) are proteins that do require reduction and S-alkylation. For each protein, we prepared a stock solution of 50 pmol/µl. Different test mixtures containing 50, 25, and 12.5 pmol/µl of each protein were prepared when needed. For routine identification in proteomic experiments, sample amounts were at least one picomol, which generally equates a clear Coomassie-stained gel band or spot. The amount of protein sample applied on the gel (25 or 50 pmol) was well above these amounts (and the detection limit of the instrument) in order to ensure that a more than adequate signal was observed while the method was being investigated.

SDS-polyacrylamide gel electrophoresis

A mixture of test proteins was electrophoresed according to Laemmli. 12% Tris-glycine gels of a thickness of 1 mm containing 10 wells were casted. Electrophoresis was carried out using a SE250 Mighty Small II apparatus (Hoefer Scientific) at room temperature. We mixed the protein samples with sample buffer (1/1, v/v) containing β-mercaptoethanol as the reducing agent and bromophenol blue to visualize the electrophoresis front. The sample was heated briefly (90-95°C, 5 min) before it was loaded on the gel. The electrophoresis running buffer was 25mM Tris base, 192mM glycine, and 0.1% SDS (w/v). Electrophoresis was carried out at 150 V for ± 1.5 hours until the dye marker had reached the edge of the gel. After fixation (2% H_3PO_4/50% ethanol/MQ; 30 minutes), we stained the proteins for ± 30 minutes with Coomassie blue G-250 at 0.2% (w/v) in 34% methanol/17% ammonium sulfate containing 3% fosforic acid. Destaining was carried out overnight using a 30% methanol solution. We excized the separated proteins and washed the gel pieces twice with 150 µl 200mM NH_4HCO_3/50% ACN for 30 minutes at 30°C, with a reswelling step in between (30 µl MQ for 10 minutes). Subsequently, the gel pieces were dried in a Speedvac (Thermo Savant) and reduced with 15 µl 10mM dithiothreitol (DTT) in 7M GuHCl/0.3M Tris, pH 9.0 (45 minutes at 55°C). Alkylation was performed by adding 5 µl 55mM iodoacetamide (IAA) and incubation in the dark for an additional 45 minutes (room temperature).

2D-PAGE gel electrophoresis

A mixture of test proteins (varying amounts, 5 to 100 pmol) was resolubilized in a rehydratation solution (6M urea, 4% (w/v) CHAPS, 75mM DTT) containing a 3-10 ampholyte stock solution (Biorad). We applied the whole mixture to an immobilized pH gradient (IPG) strips (Bio-Rad) and performed the rehydratation procedure for 6 to 8 hours at room temperature. Isoelectrofocusing (IEF) was performed at 18°C using a standard program as provided by the manufacturer. The equipment for the rehydratation and for running the IPG gels (Multiphor II) was purchased from Amersham Biosciences. After focusing, the

strips were equilibrated in 5 ml of 50mM Tris-HCl (pH 8.8)/6M urea/30% glycerol/2% SDS and 1% DTT. After 10 minutes the reducing solution was replaced by the acylating solution (same solution with DTT replaced by 2.5% IAA) for 10 minutes at room temperature. For the second dimension, we ran the 12.5% SDS-PAGE gels (18 cm x 18 cm) in a Protean II (Bio-Rad) electrophoresis apparatus at 8°C and ± 100 mAh/gel until the bromophenol blue front reached the bottom of the gel. The gels were fixated, stained with CBB and destained as described above.

Bacterial growth and preparation of extracts

S. oneidensis MR-1 was grown aerobically overnight in 20 ml Luria-Bertani (LB) medium in a rotary shaker at a speed of 200 rpm at 28°C until exponential phase (OD_{600} = ±1). The cells were centrifuged and washed twice using a 50mM Tris-HCl solution (pH 8) containing 5mM ethylenediaminotetraacetic acid (EDTA). We lysed the bacteria using 9M urea, containing 2% 3-[(cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), 1% DTT and 0.8% ampholines, after which we centrifuged the pellets at 14,000 rpm. We loaded a volume of 20 µl of bacterial extract (± 100 µg of protein as determined by a Bradford test) on a 18 cm IPG strip, pH range 4-7 (Pharmacia) via the passive in-gel rehydration protocol. 2D-PAGE was further performed as described above.

CNBr cleavage

After visualization and destaining, we excised the bands or spots containing the protein from the gels. Before cleavage with CNBr, the gel pieces were washed twice with 150 µl 50% ACN/MQ, dehydrated with 40 µl ACN, and reswollen in 5 µl milli-Q (MQ). CNBr cleavage was started by adding 5 µl 3M/ACN CNBr (Sigma) and 15 µl TFA (Applied Biosystems),(CNBr and TFA are highly toxic and corrosive products which, at any time, must be manipulated under a fumehood by skilled personnel wearing protective clothing!). After incubation overnight (4°C) the supernatant was collected and the peptides were extracted twice with 50 µl 70% ACN/0.1% TFA for 30 minutes at 37°C. All fractions were pooled and dried in a SpeedVac and redissolved in 10 µl 10mM ammonium acetate (pH 5.4) prior to carboxypeptidase treatment. Test proteins in solution (5-100 pmol) were reduced using 10 µl 10mM dithiothreitol (DTT) in 7M GuHCl/0.3M Tris, pH 9.0 (45 minutes at 55°C). Alkylation was performed by adding 10 µl 55mM iodoacetamide (IAA) in 200mM NH_4HCO_3 (pH 7) followed by incubation in the dark for an additional 45 minutes at room temperature. We performed the desalting using a Prosorb device (Applied Biosystems), the PVDF-membrane was washed twice with 100 µl water to remove salts. Following excision, the membrane was incubated in 5 µl 5.0M CNBr/ACN (Sigma) and 15 µl TFA. After incubation overnight (4°C) the supernatant was collected and the peptides were extracted and pooled as described above. Care was taken during sample manipulation to avoid oxidation. Artefactual oxidation of methionine to sulfoxide means that CNBr is inhibited from reacting with the sulfur of the methionine residue.

Carboxypeptidase digest

Carboxypeptidase Y (CPY) sequencing grade (EC 3.4.16.1) and carboxypeptidase P (CPP) sequencing grade (EC 3.4.16.1) were obtained from Roche. We redissolved the vials, containing 20 µg of enzyme, in 70 µl redistilled water, resulting in a stock solution of 5 pmol CPX/µl 40mM sodium citrate buffer (pH 6.0). Dilutions from this solution were freshly prepared. For time dependent ladder formation, we incubated the resolubilized CNBr

fragments with a mixture of carboxypeptidases Y and P (CPX) (E/S of 1/50 w/w). Sample aliquots of this reaction mixture (1 µl) were taken at 0, 1, 3, 10, 20 and 30 minutes. For concentration dependent digestions, we redissolved the CNBr fragments in 5 µl ammonium acetate buffer, and aliquots of 0.5 µl were spotted on the sample plate and incubated with 0.5 µl of 5, 1, 0.2, 0.04 and 0.008 pmol/µl CPX for 10 minutes until sample evaporation terminated the reaction.

Expression, refolding, purification and proteolytic processing of procardosin A

A bacterial expression vector containing the cDNA encoding procardosin A (pET-pCA) was introduced in an *E. coli* strain BL21(DE3). We induced expression by addition of isopropyl-1-thio-β-D-galactopyranoside (0.5mM final concentration) when OD_{600} of the cells grown at 37ºC had reached 0.6. After 3 h, cells were harvested, resuspended in 50mM Tris-HCl, 50mM NaCl (pH 7.4) and lysed by adding lysozyme (100 µg/ml). After freezing and thawing, deoxyribonuclease I (100 µg/ml) and $MgCl_2$ (100mM) was added, followed by incubation at 4ºC for 1h. The inclusion bodies were washed for 3 h with 50mM Tris-HCl, 50mM NaCl (pH 7.4), centrifuged at 10,000 x *g* for 20 minutes at 4ºC, and washed again for another 3h with 50mM Tris-HCl, 50mM NaCl (pH 7.4), 0,1% Triton X-100 (v/v). After centrifugation at 10,000 x *g* for 20 minutes at 4ºC, we dissolved the purified inclusion bodies in 8M urea, containing 100mM β-mercaptoethanol, and the protein was refolded by rapid dilution (20-fold) into 20mM Tris base, pH 8.0. The protein was concentrated in a tangential flow ultrafiltration system, Pellicon 2 (Millipore), centrifuged at 50,000 x *g* for 20 minutes at 4ºC; the supernatant was applied on a S-300 gel filtration chromatography column (Amersham Biosciences) in 20mM Tris-HCl, 0.4M urea buffer (pH 8.0). The fractions of the second peak eluting from the column are the non-aggregated forms of the protein; they were combined for further purification by ion-exchange chromatography (Resource Q column, Amersham Biosciences) eluting with the same buffer as for the S-300 column, using a 0-0.5M NaCl gradient. The purified recombinant procardosin A was activated by incubation with 0.1M sodium citrate, pH 4.0, at 37ºC. Aliquots were taken at different time points and mixed 1:1 (v/v) with SDS-sample buffer..

Spiking experiment

For the spiking experiment, 1 µl of bacterial extract (26 µg of protein) was mixed with 1 µl of the procardosin solution (0.25 µg of protein). Spiking was performed with intact procardosin A and with autoactivated protein after 8, 24 and 72 hours. The mixtures were analyzed using either 2D-PAGE or SDS-PAGE. In order to determine the position of the processed cardosin fractions in the SDS-PAGE gel, 0.25 µg of the processed fractions was analyzed in separate lanes.

Mass spectrometry

In this study, we used an Applied Biosystems 4700 Proteomics Analyzer with TOF/TOF optics (Applied Biosystems). The instrument uses a 200-Hz frequency tripled Nd:YAG laser operating at a wavelength of 355 nm. For MS/MS, ions generated by the MALDI process were accelerated at 8 kV through a grid at 6.7 kV into a short, linear, field-free drift region. In this region, the ions passed through a timed-ion-selector device that is able to select one peptide, from a mixture of peptides, for subsequent fragmentation in the collision cell. After a peptide at a given m/z was selected by the timed-ion-selector it passed through a retarding lens where the ions were decelerated and then passed into the collision

cell, which was operated at 7 kV. The collision energy is defined by the potential difference between the source and the collision cell (1 kV). After passing through the collision cell, the ions (both intact peptide ion and fragments) were accelerated in the second source region at 15 kV, passed through a second, field-free, linear drift region, into the reflector, and finally, to the detector. The detector amplifies and converts the signal to electric current, which is observed and manipulated on a PC-based operating system. For high resolution analysis, the instrument was operated in the reflector mode. After the MALDI process generates the peptide ions, they are accelerated at 20 kV through a grid at 14 kV into the first, short, linear, field-free drift region. After this point, the rest of the instrument can be treated as a continuation of this region until the ions enter the reflector and then reach the detector where, as before, the signal at the detector is amplified and converted to an electrical current.

The samples were prepared by applying 0.5 µl of the sample to a stainless steel 192-well target plate and by adding 0.25 µl matrix solution (a saturated α-cyano-4-hydroxycinnamic acid solution in 50% ACN containing 0.1% TFA). The samples were allowed to air-dry at room temperature and were then inserted into the mass spectrometer. Prior to MALDI-MS analysis, the instrument was externally calibrated with a mixture of Angiotensin I, Glu-fibrino-peptide B, ACTH (1-17), and ACTH (18-39). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrino-peptide.

Mascot analysis

For identification, a database search using a local Mascot server (Perkins *et al*, 1999) was performed against a comprehensive *Shewanella* sequence database downloaded from the Institute of Genomic Research (http://www.tigr.org).

Results

Optimization of the protocol

Regardless of the separation method (gel or solution), proteins must be prepared in a way that ensures efficient separation and that precludes undesirable artifacts. This comes down to a protocol of solubilization, denaturation, reduction and alkylation, as schematically represented (Figure 2.6). CNBr cleaves at the C-terminal side of methionine, converting it into a homoserine lactone (hsl) ($\Delta m = -48$ Da), which can undergo hydrolysis to form homoserine ($\Delta m = -29.99$ Da). Acidic conditions favour the formation of hsl, whereas under basic conditions the free acid is formed (Murphy *et al*, 1995). In this study, the best results were obtained using CNBr in 70% aqueous trifluoroacetic acid (TFA). After cleavage, the supernatant was collected and the CNBr fragments extracted.

For C-terminal sequence analysis, carboxypeptidases Y and P (CPY & CPP) are chosen most often because of their broad AA specificity (Patterson *et al*, 1995). As expected, the rate at which AA are cleaved by a mixture of CPY and CPP (CPX) depended to a great extent on the sequence of the substrate. Additionally, the cleavage rate depended on reaction conditions such as pH, ionic strength and substrate concentration. It is important to note that when a slowly released residue is followed by a rapidly released residue, this usually results in the occurrence of gaps in the sequence. Therefore, the time points at which individual samples are analyzed are critical if one wants to determine an uninterrupted sequence. Analyses were also performed by parallel microdigestions using different CPX concentrations (Patterson *et al*, 1995). An advantage here is that digestions can be performed on the probe.

Application of this technique resulted in the formation of ladders comparable to those observed in time-dependent experiments (results not shown).

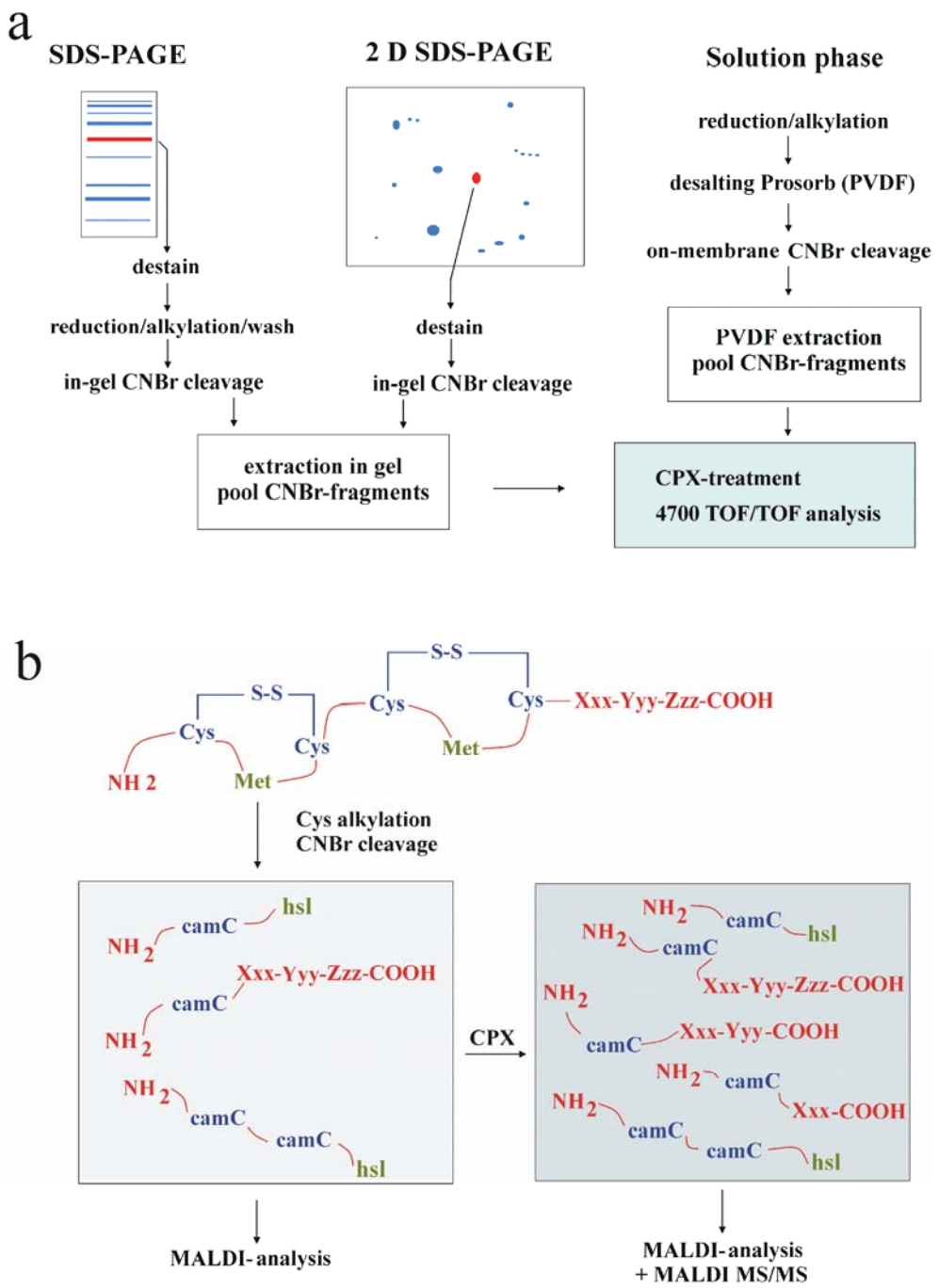


Figure 2.6. Schematic representation of the different steps in the C-terminal sequencing method**.** (**a**) Proteins separated by SDS-PAGE or 2D-PAGE were cleaved in-gel with CNBr after destaining. For SDS-PAGE separated proteins, an additional reduction, alkylation and wash step is required. Removal of excess salts was accomplished by washing the gel pieces prior to chemical cleavage. Proteins in solution were desalted on a PVDF membrane after modification, and CNBr cleavage occurred on the membrane. The resulting CNBr fragments were extracted from the gel or the PVDF, pooled and dried, and finally redissolved in buffer for incubation with CPX. (**b**) After reduction, iodoacetamide (IAA) reacts with the ionized, free –SH groups of Cys, resulting in the formation of a carboxyamidomethylcysteine (camC) derivative (Δm = +57.02 Da). CNBr cleavage results in the formation of internal fragments ending at a homoserine lactone (hsl) derivative. Only the original C-terminal sequence (Xxx-Yyy-Zzz) is accessible for enzymatic degradation (CPX) and forms a ladder that is analyzed using MALDI mass spectrometry.

C-terminal sequence analysis of model proteins

A panel of 7 standard proteins, with highly different structural properties, was chosen to demonstrate the feasibility of this approach. The mixture was separated by SDS-PAGE and 2D-PAGE and subjected to the optimized protocol. Individual standard proteins were also modified and cleaved in solution. After extraction, the unseparated peptide fragments were digested with the CPX-mixture and analyzed at different time points (Table 2.1).

As an example, the CNBr fragments from pigeon cytochrome c (PCC) were analyzed after different incubation times (Figure 2.7). Only the C-terminal peptide at 2648.32 Da (Ile81-Lys104) was degraded, whereas the internal fragment at 1762.01 Da (Glu66-Met80) remained intact. The larger N-terminal fragment (AcGly-Met65) was not observed in positive mode reflectron analysis.

Table 2.1. C-terminal sequence analysis of standard proteins

Protein	Mw(kDa)	CNBr fragment	Mw Calc(Da)[a]	Mw Obs(Da)[b]	CPX[c]
PCC	11.5	Gly1-Met65	7056.52	n.o.[d]	n.a.[e]
		Glu66-Met80	1761.90	1762.01	x
		Ile81-Lys104	2646.56	2648.32	+ (12 AA)
Ribonuclease b	14.1	Lys1-Met13	1497.74	1498.26	x
		Asp14-Met29	1660.60	n.o.	n.a.
		Lys31-Met79	5673.67	n.o.	n.a.
		Ser80-Val124	5084.39	5086.68	+ (9 AA)
MYO	16.9	Gly1-Met55	6212.15	n.o.	n.a.
		Lys56-Met131	8157.43	n.o.	n.a.
		Thr132-Gly153	2511.35	2512.26	+ (12 AA)
BLG	18.6	Leu1-Met7	756.41	n.o.	n.a.
		Lys8-Met24	1831.96	1832.8/1849.7[f]	x
		Ala25-Met107	9441.99	n.o.	n.a.
		Glu108-Met145	4283.05	4286.9/4314.9[g]	x
		His146-Ile162	2121.04	2122.9	+ (4 AA)
β-casein	24.0	Arg1-Met93	10896.2	n.o.	n.a.
		Gly94-Met102	899.48	n.o.	n.a.
		Ala103-Met109	791.40	n.o.	n.a.
		Pro110-Met144	4018.05	n.o.	n.a.
		His145-Met156	1332.67	1333.82	x
		Phe157-Met185	3187.71	3189.07	x
		Pro186-Val209	2662.53	2663.78	+ (6 AA)
YADH-1	37.1	Ser1-Met75	8176.3	n.o.	n.a.
		Gly76-Met98	2534.21	n.o.	n.a.
		Ala99-Met168	7533.5	n.o.	n.a.
		Ala169-Met193	2337.20	2354.50[f]	x
		Gly194-Met270	7947.1	n.o.	n.a.
		Pro271-Met332	6738.5	n.o.	n.a.
		Glu333-Lys347	1677.91	1665.04[h]	+ (2 AA)
BSA	68.2	Asp1-Met87	10060.8	n.o.	n.a.
		Ala88-Met184	11779.5	n.o.	n.a.
		Arg185-Met443	30327.1	n.o.	n.a.
		Pro444-Met545	11968.1	n.o.	n.a.
		Glu546-Ala581	3959.87	3962.04	+ (4 AA)

[a] Molecular weight (Mw) calculated by using the residual monoisotopic mass values with Met -> hsl, and Cys -> camC; [b] Mw observed in positive mode reflectron analysis (singly protonated); [c] x indicates that no ladder formation was observed whereas + indicates ladder formation (the number of C-terminal amino acids (AA) that was determined is indicated between brackets); [d] n.o., not observed; [e] n.a., not applicable; [f] Oxidized fragment (Δm = 16 Da); [g] fragment Glu108-Met145 from variant A (Ala>Val); [h] Conflict Ile/Val reported at position 337 in the Swiss-Prot entry (P00330). The observed mass of the C-terminal fragment, 1665.04 Da, is in agreement with a Val residue (calculated monoisotopic mass 1,663.89 Da) and was confirmed by MS/MS analysis of the C-terminal fragment. The position of the CNBr fragments in the protein sequence is indicated in arabic numbers (according to the latest Swiss-Prot release).

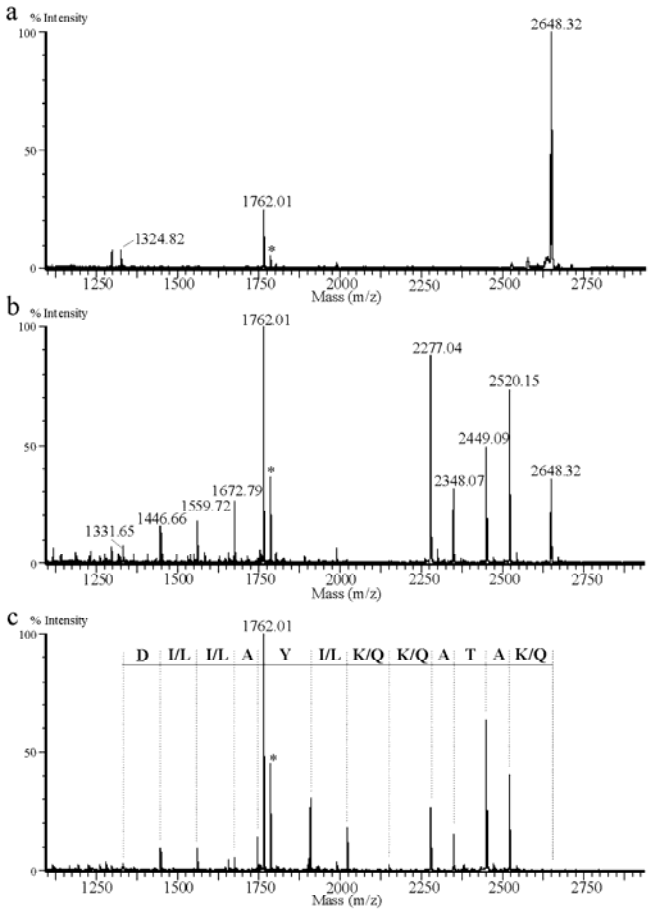


Figure 2.7. C-terminal sequence analysis of pigeon cytochrome c. After derivatization, PCC was cleaved with CNBr in solution and incubated with CPX for (a) 0, (b) 10 and (c) 20 minutes (mass spectra in positive reflectron mode, analyzed amount on probe: 1.25 pmol). Only the C-terminal peptide at 2648.32 Da formed a ladder (12 AA, indicated in one-letter code) whereas the other, internal fragment, remained intact (1762.01 Da)(Table 2.1). Sodium adducts are indicated with *.

Proteome analysis of Shewanella oneidensis MR-1

The method was applied for proteome analysis of proteins from *Shewanella oneidensis* MR-1. Since this bacterium is able to reduce a variety of metal substrates, it is of considerable interest to researchers involved in bioremediation (Glasauer *et al*, 2002). Although a limited number of proteomic studies of *Shewanella* have been reported (Vanrobaeys *et al*, 2003), only a fraction of the *Shewanella* genes have been characterized at the protein level so far.

The total protein extract from aerobically grown *Shewanella* MR-1 was separated by 2D-PAGE. After Coomassie staining, 25 spots were randomly selected and subjected to the new protocol (Figure 2.8). Of the 25 spots, 19 proteins with a Mw ranging from 10 to 57 kDa could be identified by PMF-analysis of the CNBr fragments and by performing a database search on a local MASCOT server (Perkins *et al*, 1999). Upon incubation with CPX, we observed ladder formation for 11 proteins and one to ten AA were cleaved from the C-terminal peptide (Table 2.2).

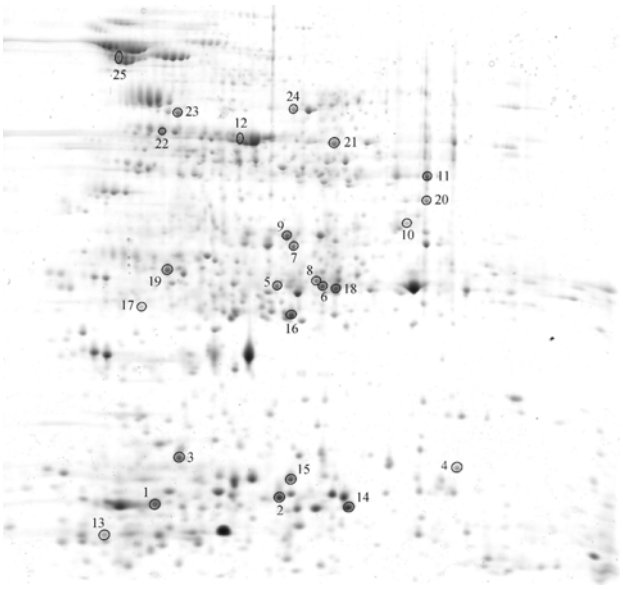


Figure 2.8. 2D-PAGE separated proteins from aerobically grown Shewanella MR-1. +/- 100 µg of total cell extract was loaded on an IPG (4-7) strip and analyzed as decribed. The labeled spots were subjected to the new approach.

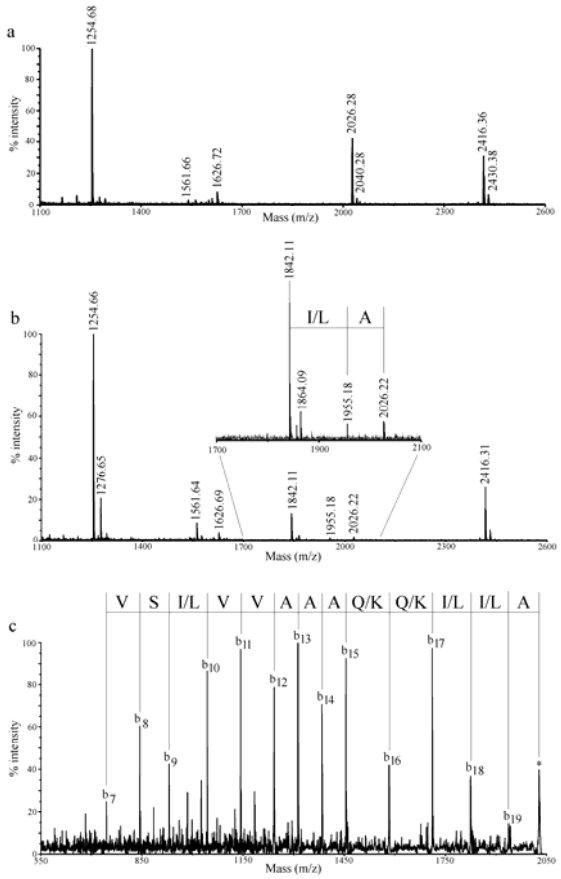


Figure 2.9. C-terminal sequence analysis of 2D-PAGE separated proteins from *Shewanella* MR-1**.** (a) Mass spectrum in the positive reflectron mode of CNBr fragments from the gel spot containing uridine phosphorylase. (b) Upon incubation with CPX only two C-terminal amino acids were removed from the C-terminal fragment at m/z 2026.28 (Lys233-Ala252, NCBI *Entrez* [GI_24375619]). (c) MALDI MS/MS fragmentation of the C-terminal fragment (precursor labeled with an asterisk) resulted in a continuous stretch of b-ions from which 13 C-terminal amino acids were identified.

In cases where only a limited amount of sequence information was obtained, the C-terminal CNBr fragment was selected for MS/MS analysis. This approach is demonstrated using the CNBr peptide map of uridine phosphorylase of which, upon incubation with CPX, only two C-terminal AA were removed (Figure 2.9). The peptide at m/z 2026.28 was selected as precursor and yielded, upon fragmentation, a continuous stretch of b-ions from which 13 C-terminal AA were identified. Addition of the C-terminal sequence as a search constraint in MASCOT significantly improved the identification score. Furthermore, it allowed the identification of two more proteins of which only one CNBr fragment (the C-terminal fragment) was observed (Table 2.2a). Although four CNBr fragments were observed, cleavage of the translation elongation factor Ts resulted in a C-terminal fragment of 5430.9 Da, too large to be measured in reflectron analysis mode. Database analysis of the other identified proteins of which the C terminus was not characterized (ten spots) indicated that all of them, except two, had a C-terminal peptide with a Mw below 1 kDa or higher than 5 kDa (Table 2.2b). Peptide fragments with such Mw cannot be analyzed with sufficient resolution in reflectron analysis at this sensitivity. The fact that we did not observe ladder formation for spots 15 and 24, both having a C-terminal peptide with a Mw in the correct range, indicated that these fragments were not efficiently extracted or did not ionize well.

Table 2.2a. C-terminal sequence analysis of 2D-PAGE separated proteins of *Shewanella oneidensis*

Protein[a]	Spot[g]	Mw (kDa)	Mw Obs [b] (Da)	CNBr fragment	Mw Calc [c] (Da)	CPX-sequence [d]
[GI_24371821][NP_715863]	1	12.5	1489.81	Ser2-Met15	1488.76	x
Ribosomal protein L7/L12			3495.93	Ser89-Lys122	3494.87	+ (EIK)
[GI_24375179][NP_719222]	2	15.6	1367.86	Ile51-Met63	1366.74	x
Universal stress protein family			2673.63	Ala27-Met50	2672.38	x
			3667.29	Val109-Lys143	3665.95	+ (AVCPVLVVK)
[GI_24372148][NP_716190]	3	19.7	1288.68	Ser145-Met156	1287.62	x
Hypothetical protein S00554			1514.93	Leu162-Leu174	1497.89[e]	+ (KL)
			2470.37	Arg97-Met118	2469.29	x
			2537.33	Ser119-Met141	2536.24	x
[GI_24374881][NP_718924]	4	20.4	2951.66	Lys165-Lys191	2950.61	+ (RK)
Conserved hypothetical protein						
[GI_24375619][NP_719662]	5	27.1	1254.68	Ala2-Met13	1253.65	x
Uridine phosphorylase			2026.28	Lys233-Ala252	2025.26	+ (LA)
			2416.36	Leu14-Met36	2415.32	x
[GI_24376191][NP_720235]	6	29.3	2090.04	Ile257-Glu274	2089.04	+ (FKATYSE)
Conserved hypothetical protein			2339.25	Thr85-Met107	2338.14	x
[GI_24373198][NP_717241]	7	30.5	1254.68	Asn201-Met212	1253.60	x
Translation elongation factor Ts			1738.95	Lys218-Met233	1737.90	x
			1840.02	Ala2-Met19	1838.97	x
			3457.69	His169-Met200	3456.73	x
[GI_24371943][NP_715985]	8	31.9	2015.48	Val231-Met249	2014.14	x
Methylisocitrate lyase			2083.48	Val136-Met154	2082.12	x
			2280.46	Thr210-Met230	2279.07	x
			2637.73	Glu113-Met135	2636.27	x
			2771.91	Ala155-Met182	2770.42	x
			3070.18	Ile183-Met209	3068.64	x
			3293.19	Gln266-Lys292	3291.62	+ (DK)
			3451.42	Met1-Met32	3449.82[f]	x
[GI_24372359][NP_716401]	9	32.3	1761.12	Leu296-Lys311	1760.01	+ (FVK)
Malate dehydrogenase						
[GI_24374179][NP_718222]	10	37.3	1041.58	Ile125-Met133	1040.47	x
Leucine dehydrogenase			1062.70	Ala335-Ala344	1061.60	+ (AKA)
			2337.26	Trp45-Met65	2320.08[e]	x

[GI_24346525][Aan54007] Fructose-bisphosphate aldolase II	11	38.8	1680.95 3037.39 3272.52 3730.61	Tyr341-Leu355 Glu313-Met340 Ala2-Met31 Thr67-Met99	1679.89 3036.49 3271.65 3729.73	+ (L) x x x
[GI_416942][P33169] Elongation factor Tu	12	43.5	2532.63 2555.45	Pro114-Met135 Asp370-Ala394	2531.41 2554.43	x +(GAGVVAKIIA)

Table 2.2b. Proteins of *Shewanella oneidensis* identified by PMF-analysis of CNBr fragments without characterization of the C termini

Protein[a]	Spot[g]	Mw (kDa)	Calc Mw of C-terminal peptide (Da)[c]
not identified	13	-	-
not identified	14	-	-
[GI_32171551] [Q8EAH2] 30S ribosomal protein S6	15	15.0	3206.5
[GI_24372802] [NP_716844.1] Purine nucleoside phosphorylase	16	25.6	817.47
[GI_24373389] [NP_717432.1] Hypothetical protein SO1825	17	25.2	5185.78
[GI_24375475] [NP_719518.1] Aerobic respiration control ArcA	18	27.2	5687.93
not identified	19	-	-
[GI_24347242] [AAN54551.1] Alcohol dehydrogenase II	20	40.0	7147.7
not identified	21	-	-
[GI_30315823] [Q8EBR0] 2-phosphoglycerate dehydratase	22	45.6	7843.13
[GI_24373359][NP_717402.1] Trigger factor	23	47.6	717.4
[GI_24376221] [NP_720265.1] ATP synthase F1, alpha subunit	24	55.1	4189.11
[GI_24372295] [NP_716337.1] Chaperonin GroEL	25	57.0	150.06

[a] NCBI *Entrez* entries (http://www.ncbi.nih.gov/Entrez/); [b] Mw observed in positive mode reflectron analysis (singly protonated); [c] Mw calculated by using the residual monoisotopic values with Met -> hsl, and Cys -> camC; [d] x indicates that no ladder formation was observed whereas + indicates ladder formation (the observed amino acid sequence is indicated from N -> C); [e] CNBr fragment containing an oxidized Trp (Δm = 16 Da); [f] indicates a CNBr fragment with a missed Met-Xxx cleavage. [g] Spot number according to the position on the 2D-PAGE (Figure. 2.8). The position of the CNBr fragments in the protein sequence is indicated in arabic numbers (numbering according to the NCBI entries).

Proteolytic processing of procardosin A

Cardosin A is an aspartic proteinase that is synthesized as a single chain precursor and which undergoes proteolytic processing to generate the active two-chain mature enzyme. During this process, an N-terminal prosegment and an internal segment, known as PSI (plant-specific insert), are excised from the precursor (Simoes *et al*, 2004). To characterize in detail the proteolytic processing, a recombinant precursor was produced in *E.coli* and activated by incubation at pH 4.0 at 37°C. The processing was followed by SDS-PAGE analysis (Figure 2.10a). Cleavage of the prosegment is known to result in the formation of two fragments, B1 and B2, of which the N and C termini were determined by Edman degradation and chemical C-terminal sequence analysis. This indicated that cleavage occurred at the Phe63-Arg64 peptide bond. The B1 fragment had the same C terminus as the intact precursor, indicating that no processing had taken place, and was processed in two fragments, C1 and C2, which

finally rendered the active mature enzymes D1 and D2 (Figure 2.10b). After electroblotting and staining, the smallest fragments (C2 and D2) were no longer observed on the PVDF-membrane. Although the larger fragments (C1 and D1) were slightly visible, the amounts present on the membrane were insufficient to obtain N- or C-terminal sequence information by chemical sequence analysis. Therefore, our method was applied on the SDS-PAGE separated fragments.

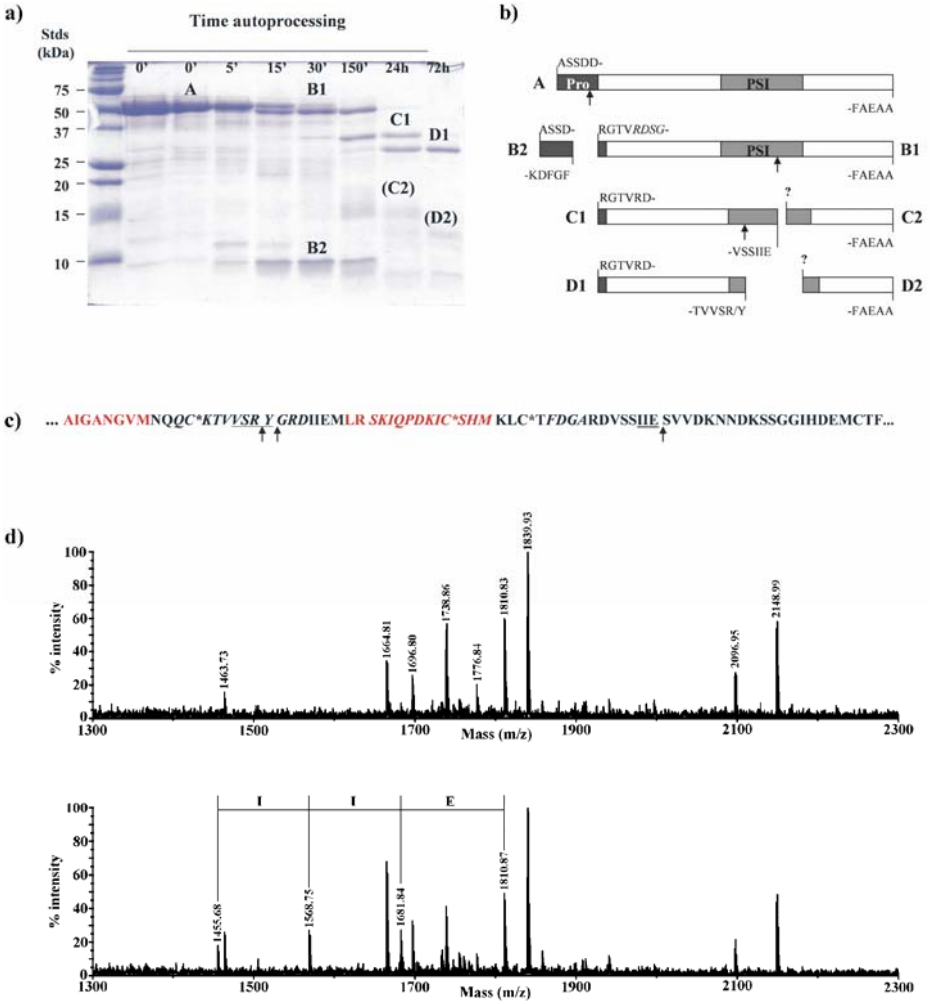


Figure 2.10. (**a**) Proteolytic processing of recombinant procardosin A analyzed by SDS-PAGE. The precursor is labeled above the band (A) and the processed forms are labeled as shown. (**b**) Cleavage of the prosegment (Pro) occurred 5 amino acids closer to the N terminus than in the native protein (indicated in italic). B1 is further processed by removal of the plant specific insert (PSI). MALDI analysis of the CNBr peptides of fragments C2 and D2 indicated the presence of the C-terminal peptide of the intact protein. (**c**) AA sequence of the PSI domain in which the processing occurred (TrEMBL Q9XFX3). The arrows indicate the C-terminal processing sites. Fragment C1 ends on IIE, whereas the D1 fragment has a ragged C terminus, VSR/Y. C-terminal AA identified upon CPX incubation are underlined whereas AA determined by MS/MS analysis are indicated in italic. Other observed CNBr peptides are indicated repetitively in black and red (C* = camC derivative). (**d**) C-terminal sequence analysis of SDS-PAGE separated proteins from *Shewanella* spiked with the processed (24h) procardosin A. The upper panel indicates the CNBr peptide map of the proteins present in gel spot two (Figure 2.11). The lower panel shows the same mass spectrum upon incubation with CPX. The fragment at m/z 1810.63 is the C-terminal peptide of fragment C1. Two peptides (m/z 1,664.81 and 2,148.99) are internal fragments of C1. We also observed three internal fragments of the *Shewanella* translation elongation factor Ts (m/z 1463.73, 1738.86 and 1839.93) (Table 2.2a).

Upon CNBr cleavage of C1 in the gel spot, three peptide fragments were observed during MALDI analysis. Upon incubation with CPX, only one fragment (1810.75 Da, 1

camC) showed ladder formation. MS/MS analysis of the fragment yielded the sequence XFDGAX, indicating that fragment Lys345-Glu360 is the C-terminal fragment. As expected from previous studies, the first processing step occurred somewhere in the middle of the PSI-domain (Ramalho-Santos *et al*, 1998). MS/MS analysis of the other fragments indicated that these were internal CNBr-peptides preceding the C-terminal fragment in the PSI-sequence (Figure 2.10c). Application of the procedure on the D1 gel spot resulted in a less complex mass spectrum, as only two peptide fragments were observed. Upon incubation with CPX, both fragments (m/z 1219.55 and 1382.62) showed the same ladder formation (…VSR/Y) (Figure 2.10c). This indicated that both fragments are the same C-terminal peptide, having a ragged C terminus R/Y (Δm = 163.06 Da). After CNBr cleavage, MALDI analysis of the less intense C2 and D2 bands indicated the presence of a fragment at 2240.23 Da. Although the low intensity of the observed peak excluded incubation with CPX, this mass was in perfect agreement with the mass of the C-terminal CNBr fragment of the intact protein (2239.11 Da), indicating that processing does not occur at the C termini of these fragments.

C-terminal sequence analysis in complex sample mixtures

In a final experiment, we tested whether the method can be used to identify the proteolytic processing of a protein in a complex sample mixture. Therefore, the total *Shewanella* protein extract was spiked with the intact and the processed recombinant procardosin A (the latter at a concentration of 1% of the total amount of protein). The mixtures were first analyzed using a 2D-PAGE differential display approach. Application of the method indicated the C-terminal sequence of the intact protein as well as the C termini of fragments C1 and D1 (results not shown).

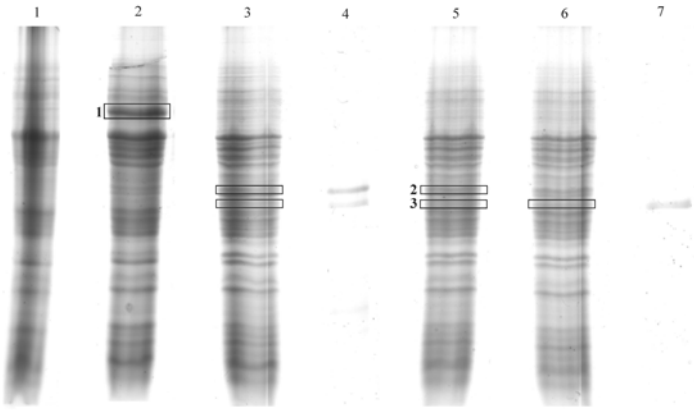


Figure 2.11. SDS-PAGE of the spiked *Shewanella oneidensis* protein extract. Lanes 1 and 2 contained respectively 26 µg of the total *Shewanella* extract without and with 0.25 µg procardosin A. The *Shewanella* extract was spiked with procardosin after after 8 (lane 3), 24 (lane 5) and 72 hours (lane 6). The processed forms were separated individually after 24 (lane 4) and 72 hours (lane 7) of autoactivation at 37°C. Gel regions that were cut C-terminal sequence analysis are boxed.

A spiking experiment with 1% of the sample being recombinant procardosin A was performed, to prove the applicability of the new approach on complex samples. 2D-gels with either *Shewanella* extract alone or *Shewanella* extract spiked with processed procardosin A after respectively 0, 8, 24 and 72 hours of autoactivation were run and the gel images compared. The differentially displayed spot were excised and the C-terminus determined (results not shown). To further increase the complexity of the samples, the mixtures were also analyzed by 1D SDS-PAGE. The intact procardosin A was clearly observed in the spiked

Shewanella extract (Figure 2.11). For gel spot one, we observed, upon CNBr cleavage, one dominant peak (m/z 2240.16) degrading upon CPX incubation. MS/MS analysis confirmed that this was the C-terminal peptide of the intact protein (Figure 2.12). After 8 hours of activation the formation of the C1 fragment was observed. Upon 24 hours of activation, both fragments C1 and D1 appeared, whereas upon 72 hours of activation, only fragment D1 was observed. The method was applied on the gel pieces in the spiked *Shewanella* samples after 24 hours of autoactivation (Figure 2.11, gel spots 2 & 3). As illustrated, this resulted in a much more complex mass spectrum (Figure 2.10d). Upon CPX-incubation, only one fragment (m/z 1810.87) indicated ladder formation. The observed C-terminal sequence, Ile-Ile-Glu, confirmed that this fragment is the C-terminal peptide of the processed C1 protein. Similarly, C-terminal sequence analysis of the complex mixture observed in spot three indicated that this fraction contained the C-terminal peptide of fragment D1 (m/z 1382.64 Da) (results not shown).

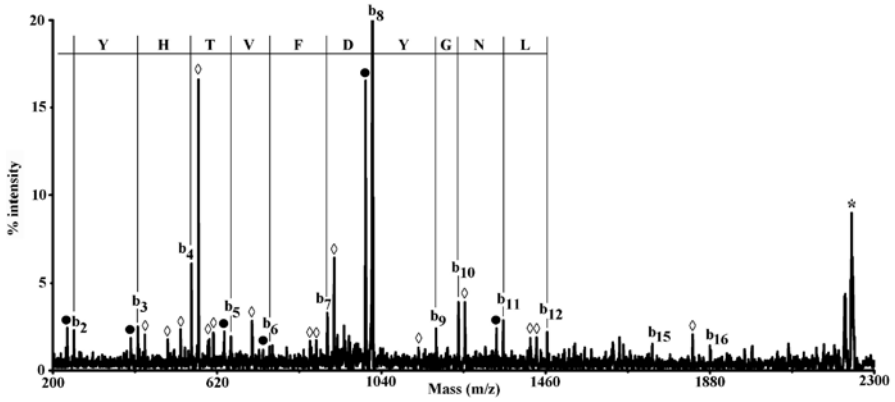


Figure 2.12. C-terminal sequence analysis of the *Shewanella* extract spiked with the intact procardosin protein. MALDI MS/MS fragmentation of the C-terminal fragment of the intact procardosin protein spiked in the *Shewanella* extract (precursor m/z 2240.16 labeled with *) A series of b-ions was observed from which 10 amino acids were determined. Filled circles indicate (b-17)-ions whereas open diamonds are internal fragment ions or a-ions.

Discussion

We described a novel approach that allows the systematic identification of the C termini of proteins. First, we demonstrated that the method is compatible with sample preparation techniques used in proteome analysis. Depending on the purification strategy chosen, the total number of manipulations is limited to 3 or 4 (Figure 2.6), which is equal or less than the number of steps involved in enrichment strategies reported for the study of e.g. protein phosphorylation (Oda *et al*, 2001; Zhou *et al*, 2001). Longer incubation times with CPX may result in a decrease in ion intensity of some of the fragments (Figure 2.7c). Therefore, in order to obtain a complete sequence, without gaps, it is necessary to analyze the ladder fragments at different time points (Patterson *et al*, 1995).

A positive identification of the C terminus will depend on the length and ionization capacity of the generated CNBr fragments. In our approach, MS analysis was performed using MALDI ionization which generates predominantly singly charged ions. The use of the α-cyano-4-hydroxycinnamic acid matrix produces interference in the low Mw mass range and therefore presents a challenge for the analysis of peptides with a Mw below 0.7-1.3 kDa. In our experience, the upper mass limit for the analysis of ladder sequences in RE-MALDI-MS,

providing enough resolution and accuracy to identify AA by 0.1 Da weight difference, is restricted to C-terminal fragments with a Mw of ± 5 kDa. In *Shewanella*, methionine residues occur at a frequency of 2.6% (http://www.ebi.ac.uk/proteome/SHEON). Statistical analysis of the *Shewanella* protein database indicated that, for 3432 proteins, a methionine occurs within the region of the last 50 residues. In 2559 proteins the methionine residue does not occur in the last 10 residues. This indicated that approximately 50% of all theoretical *Shewanella* proteins (5,177 ORF's) have a C-terminal fragment that is identifiable using this approach. The use of other MS techniques might allow the identification of larger C-terminal fragments. Recently, it has been demonstrated that ions in the 5000-10000 m/z range can be analyzed with errors less than 27 ppm using MALDI-FTMS (Jones *et al*, 2003).

The use of 2D-PAGE has a number of significant drawbacks. Conceptually, the simplest approach to analyze complex polypeptide mixtures in a gel-free system is obtained by using the multidimensional protein identification technology (MudPIT) (Washburn *et al*, 2001). A gel-free approach wherein intact proteins, rather than peptides, are separated by multi-LC may be used, as our method is applicable to proteins separated in solution. Although initial reports on such separations have been published, these methods still deal with chromatographic and other difficulties (Meng *et al*, 2002).

It should be pointed out that, compared to in-gel trypsin digestion; the CNBr-approach is less sensitive. Nevertheless, sub-microgram detection is adequate for most proteomics applications and is comparable to the detection limit of Coomassie stained gel spots. Furthermore, it should be noted that most exopeptidases have a K_m-value in the range of 5-50 mM, which means that they are operating at 50% maximum velocity when a protein concentration of 5 pmol/µl is used (Quadroni *et al*, 1999).

Over the past decade, miniaturization has become important in the design of analytical devices for high-throughput applications (Mitchell, 2001). The use of microfluidic instruments offers several advantages, and MS is emerging as a detection device for such systems (Oleschuk *et al*, 2000). A microfluidic device was recently described for desalting and the enrichment of samples for MALDI analysis (Astorga-Wells *et al*, 2003). It was also demonstrated that C-terminal sequence analysis can be performed using CPY directly on a protein chip and, for two test peptides, a C-terminal sequence was obtained (Caputo *et al*, 2003). In this regard, it is feasible that our method can be performed in a microfluidic device allowing a multiplexed approach.

We have provided evidence that the C-terminal sequence, together with the Mw of the CNBr-fragments, is sufficient to identify and characterize proteins. 2D-PAGE separated proteins from *Shewanella oneidensis* were chosen as a model system to investigate the effectiveness of such an approach. A sequence tag of one to 12 C-terminal AA strongly increased the confidence level of a database 'hit' (Table 2.3). It has previously been shown that the use of a C-terminal tag is sufficient to identify proteins in small databases (Wilkins *et al*, 1998).

The method we have presented here is robust and allows the identification of the C-terminal sequence of proteins separated by SDS-PAGE, 2D-PAGE, or purified in solution. Application of this method to complex mixtures verified its vital role for the determination of C-terminal sequences of proteins at a proteomic scale. In its current form, the method is not yet suitable for high-throughput analysis, and the use of the MALDI-TOF/TOF instrument requires that the Mw of the C-terminal fragment is below 5 kDa. Further improvements will

depend on automation using a chip-based approach and on the use of other MS devices, allowing the analysis of larger C-terminal fragments.

Table 2.3. Mascot identification scores [a]

Protein [b]	Peptide mass fingerprint		PMF + C-terminal sequence tag	
Ribosomal protein L7/L12	+ [c]	(56)	+ [c]	(94)
Universal stress protein family	+	(79)	+	(191)
Hypothetical protein S00554	+	(97)	+	(119)
Conserved hypothetical protein	–		+	(53)
Uridine phosphorylase	+	(45)	+	(66)
Conserved hypothetical protein	+	(45)	+	(143)
Translation elongation factor Ts	+	(99)	not applicable	
Methylisocitrate lyase	+	(147)	+	(163)
Malate dehydrogenase	–		+	(54)
Leucine dehydrogenase	+	(78)	+	(113)
Fructose-bisphosphate aldolase II	+	(97)	+	(108)
Elongation factor Tu	+	(48)	+	(178)

[a] In MASCOT, the score is based on the absolute probability (P) that the observed match between the experimental data and the database sequence is a random event. The reported score is $-10*\mathrm{Log}_{10}(P)$. If 1.5×10^5 peptides during a search are within the mass tolerance window of the precursor mass, and the significance threshold is chosen to be 0.05 (5% chance of a false positive), this translates into a threshold score of 65. [b] NCBI *Entrez* entries. [c] + (–) indicates a positive (negative) MASCOT identification.

Cardosin A is an aspartic protease, a group of proteases that have been implicated in a variety of physiological processes where cell death events play a key role (Mutlu *et al*, 1998; Runeberg-Roos *et al*, 1998; Lindholm *et al*, 2000; Simoes *et al*, 2004). Often, the activation of these proteases triggers the onset of the events that ultimately determine the fate of the plant cell. In the particular case of cardosin A, the major milk-clotting enzymes of the flowers of cardoon and a model plant aspartic protease, characterization of the activation process has been hampered by the difficulty in isolating the precursor form from its natural source. Therefore, production of milligram amounts of recombinant procardosin A was important, not only for structural studies, but also to study its activation and proteolytic processing in more detail.

Using our approach, we demonstrated that the activation of cardosin A is a multistep process (Castanheira *et al*, 2005). The data clearly indicated that the first step is the removal of the propeptide, generating an active intermediate with the internal PSI (Plant Specific Insert) still present. This suggests that the recombinant cardosin A is inactivated by the presence of the propeptide, a method of inactivation often seen in other aspartic acid proteases (Bernstein *et al*, 1999). Furthermore, these data suggest that the presence of the PSI has little effect on the activity. Experiments using active site mutants and protease inhibitors further indicate that the activation is an autocatalytic process. The second step in the proteolytic processing of cardosin A is the removal of the PSI, with an initial cleavage in the middle of the PSI and further cleavage bidirectionally from that point towards the sequence boundaries between the PSI and the two polypeptide chains of mature cardosin A. In vitro activation of recombinant procardosin A ultimately results in the generation of an active form with the two polypeptide chains still associated by a disulfide bond. This has also been reported for cyprosin and the sunflower seed aspartic protease (White *et al*, 1999; Park *et al*, 2001). The incomplete removal of the PSI described here suggests that in vivo completion of the maturation might require the action of other protease/exopeptidase(s). Nevertheless, the production of an active form of cardosin A requires only the removal of the propeptide, which can be accomplished through autoactivation under acidic conditions. Most likely this occurs inside the vacuole at a slow rate and is accelerated by sudden decreases in pH.

References

Aebersold, R.; Mann, M. (**2003**). Mass spectrometry-based proteomics. *Nature* **422**(6928): 198-207.

Ambler, R.P. (**1965**). The behaviour of peptides formed by cyanogen bromide cleavage of proteins. *Biochem J* **96**: 32.

Astorga-Wells, J.; Jornvall, H.; Bergman, T. (**2003**). A microfluidic electrocapture device in sample preparation for protein analysis by MALDI mass spectrometry. *Anal Chem* **75**(19): 5213-9.

Bailey, J.M.; Nikfarjam, F.; Shenoy, N.R.*; et al.* (**1992**). Automated carboxy-terminal sequence analysis of peptides and proteins using diphenyl phosphoroisothiocyanatidate. *Protein Sci* **1**(12): 1622-33.

Bailey, J.M.; Shively, J.E. (**1994**). A chemical method for the C-terminal sequence analysis of proteins. *Methods: A companion to Methods in Enzymology* **6**: 334-350.

Bailey, J.M. (**1995**). Chemical methods of protein sequence analysis. *J Chromatogr A* **705**(1): 47-65.

Bailey, J.M.; Tu, O.; Issai, G.*; et al.* (**1995**). Automated carboxy-terminal sequence analysis of polypeptides containing C-terminal proline. *Anal Biochem* **224**(2): 588-96.

Bech, L.M.; Breddam, K. (**1989**). Inactivation of carboxypeptidase Y by mutational removal of the putative essential histidyl residue. *Carlsberg Res Commun* **54**(5): 165-71.

Bergman, T. **(2000)**. Ladder sequencing. ed. Jolles, P. Jornvall, H. *Proteomics in functional genomics.* Basel/Switzerland, Birkhaüser Verlag. **88:** 133-44.

Bernstein, N.K.; James, M.N. (**1999**). Novel ways to prevent proteolysis - prophytepsin and proplasmepsin II. *Curr Opin Struct Biol* **9**(6): 684-9.

Bonetto, V.; Bergman, A.C.; Jornvall, H.*; et al.* (**1997**). C-terminal sequence determination of modified peptides by MALDI MS. *J Protein Chem* **16**(5): 371-4.

Boyd, V.L.; Bozzini, M.; Zon, G.*; et al.* (**1992**). Sequencing of peptides and proteins from the carboxy terminus. *Anal Biochem* **206**(2): 344-52.

Breddam, K. (**1986**). Serine carboxypeptidases: a review. *Carlberg Res Comm* **51**: 83-128.

Brivio, M.; Fokkens, R.H.; Verboom, W.*; et al.* (**2002**). Integrated microfluidic system enabling (bio)chemical reactions with on-line MALDI-TOF mass spectrometry. *Anal Chem* **74**(16): 3972-6.

Caputo, E.; Moharram, R.; Martin, B.M. (**2003**). Methods for on-chip protein analysis. *Anal Biochem* **321**(1): 116-24.

Castanheira, P.; Samyn, B.; Sergeant, K.*; et al.* (**2005**). Activation, proteolytic processing, and peptide specificity of recombinant cardosin A. *J Biol Chem* **280**(13): 13047-54.

Chait, B.T.; Wang, R.; Beavis, R.C.*; et al.* (**1993**). Protein ladder sequencing. *Science* **262**(5130): 89-92.

Checler, F.; Vincent, B. (**2002**). Alzheimer's and prion diseases: distinct pathologies, common proteolytic denominators. *Trends Neurosci* **25**(12): 616-20.

Chou, K.C. (**2001**). Prediction of protein signal sequences and their cleavage sites. *Proteins* **42**(1): 136-9.

Chung, J.J.; Shikano, S.; Hanyu, Y.*; et al.* (**2002**). Functional diversity of protein C-termini: more than zipcoding? *Trends Cell Biol* **12**(3): 146-50.

Chung, J.J.; Yang, H.; Li, M. (**2003**). Genome-wide analyses of carboxyl-terminal sequences. *Mol Cell Proteomics* **2**(3): 173-81.

Ciurli, S.; Safarov, N.; Miletti, S.*; et al.* (**2002**). Molecular characterization of Bacillus pasteurii UreE, a metal-binding chaperone for the assembly of the urease active site. *J Biol Inorg Chem* **7**(6): 623-31.

Compagnini, A.; Cunsolo, V.; Foti, S.*; et al.* (**2001**). Improved accuracy in the matrix-assisted laser desorption/ionization-mass spectrometry determination of the molecular mass of cyanogen bromide fragments of proteins by post-cleavage reaction with tris(hydroxymethyl)aminomethane. *Proteomics* **1**(8): 967-74.

Cool, D.R.; Hardiman, A. (**2004**). C-terminal sequencing of peptide hormones using carboxypeptidase Y and SELDI-TOF mass spectrometry. *Biotechniques* **36**(1): 32-4.

Cordoba, O.L.; Linskens, S.B.; Dacci, E.*; et al.* (**1997**). 'In gel' cleavage with cyanogen bromide for protein internal sequencing. *J Biochem Biophys Methods* **35**(1): 1-10.

Dove, A. (**2001**). The bittersweet promise of glycobiology. *Nat Biotechnol* **19**(10): 913-7.

Doyle, D.A.; Lee, A.; Lewis, J.*; et al.* (**1996**). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85**(7): 1067-76.

Du, Y.; Meng, F.; Patrie, S.M.*; et al.* (**2004**). Improved molecular weight-based processing of intact proteins for interrogation by quadrupole-enhanced FT MS/MS. *J Proteome Res* **3**(4): 801-6.

Dukan, S.; Turlin, E.; Biville, F.*; et al.* (**1998**). Coupling 2D SDS-PAGE with CNBr cleavage and MALDI-TOFMS: a strategy applied to the identification of proteins induced by a hypochlorous acid stress in Escherichia coli. *Anal Chem* **70**(20): 4433-40.

Endrizzi, J.A.; Breddam, K.; Remington, S.J. (**1994**). 2.8-A structure of yeast serine carboxypeptidase. *Biochemistry* **33**(37): 11106-20.

Fanning, A.S.; Anderson, J.M. (**1999**). PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. *J Clin Invest* **103**(6): 767-72.

Gatto, G.J., Jr.; Berg, J.M. (**2003**). Nonrandom tripeptide sequence distributions at protein carboxyl termini. *Genome Res* **13**(4): 617-23.

Glasauer, S.; Langley, S.; Beveridge, T.J. (**2002**). Intracellular iron minerals in a dissimilatory iron-reducing bacterium. *Science* **295**(5552): 117-9.

Goodlett, D.R.; Armstrong, F.B.; Creech, R.J.*; et al.* (**1990**). Formylated peptides from cyanogen bromide digests identified by fast atom bombardment mass spectrometry. *Anal Biochem* **186**(1): 116-20.

Gross, E.; Witkop, B. (**1962**). Nonenzymatic cleavage of peptide bonds: the methionine residues in bovine pancreatic ribonuclease. *J Biol Chem* **237**: 1856-60.

Hardeman, K.; Samyn, B.; Van der Eycken, J.*; et al.* (**1998**). An improved chemical approach toward the C-terminal sequence analysis of proteins containing all natural amino acids. *Protein Sci* **7**(7): 1593-602.

Hardy, J.; Selkoe, D.J. (**2002**). The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297**(5580): 353-6.

Henry, C. (**1998**). Automated protein sequencers - alive and kicking. *Anal Chem* **70**: 401A-4A.

Inglis, A.S. (**1991**). Chemical procedures for C-terminal sequencing of peptides and proteins. *Anal Biochem* **195**(2): 183-96.

Ishii, S.; Yokosawa, H.; Kumazaki, T.*; et al.* (**1983**). Immobilized anhydrotrypsin as a specific affinity adsorbent for tryptic peptides. *Methods Enzymol* **91**: 378-83.

Jarrett, J.T.; Lansbury, P.T., Jr. (**1993**). Seeding "one-dimensional crystallization" of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie? *Cell* **73**(6): 1055-8.

Johnson, T.B.; Nicolet, B.H. (**1911**). Hydantoins: the synthesis of 2-thiohydantoins. *J Am Chem Soc* **33**: 1706-14.

Jones, J.J.; Stump, M.J.; Fleming, R.C.*; et al.* (**2003**). Investigation of MALDI-TOF and FT-MS techniques for analysis of Escherichia coli whole cells. *Anal Chem* **75**(6): 1340-7.

Jurgens, G. (**2004**). Membrane trafficking in plants. *Annu Rev Cell Dev Biol* **20**: 481-504.

Kaiser, R.; Metzka, L. (**1999**). Enhancement of cyanogen bromide cleavage yields for methionyl-serine and methionyl-threonine peptide bonds. *Anal Biochem* **266**(1): 1-8.

Klarskov, I.; Breddam, K.; Roepstorff, P. (**1989**). C-terminal sequence determination of peptides degraded with carboxypeptidases of different specificities and analyzed by 252-Cf plasma desorption mass spectrometry. *Anal Biochem* **180**(1): 28-37.

Knight, Z.A.; Schilling, B.; Row, R.H.*; et al.* (**2003**). Phosphospecific proteolysis for mapping sites of protein phosphorylation. *Nat Biotechnol* **21**(9): 1047-54.

Kosaka, T.; Takazawa, T.; Nakamura, T. (**2000**). Identification and C-terminal characterization of proteins from two-dimensional polyacrylamide gels by a combination of isotopic labeling and nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* **72**(6): 1179-85.

Krimmer, T.; Geissler, A.; Pfanner, N.*; et al.* (**2001**). Sorting of preproteins into mitochondria. *Chembiochem* **2**(7-8): 505-12.

Kumazaki, T.; Nakako, T.; Arisaka, F.*; et al.* (**1986**). A novel method for selective isolation of C-terminal peptides from tryptic digests of proteins by immobilized anhydrotrypsin: application to structural analyses of the tail sheath and tube proteins from bacteriophage T4. *Proteins* **1**(1): 100-7.

Lin, M.; Campbell, J.M.; Mueller, D.R.*; et al.* (**2003**). Intact protein analysis by matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **17**(16): 1809-14.

Lindholm, P.; Kuittinen, T.; Sorri, O.*; et al.* (**2000**). Glycosylation of phytepsin and expression of dad1, dad2 and ost1 during onset of cell death in germinating barley scutella. *Mech Dev* **93**(1-2): 169-73.

Loo, R.R.O.; Stevenson, T.I.; Mitchell, C.*; et al.* (**1996**). Mass spectrometry of proteins directly from polyacrylamide gels. *Anal Chem* **68**(11): 1910-17.

Mann, M.; Jensen, O.N. (**2003**). Proteomic analysis of post-translational modifications. *Nat Biotechnol* **21**(3): 255-61.

Meng, F.; Cargile, B.J.; Miller, L.M.*; et al.* (**2001**). Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat Biotechnol* **19**(10): 952-7.

Meng, F.; Cargile, B.J.; Patrie, S.M.*; et al.* (**2002**). Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Anal Chem* **74**(13): 2923-9.

Miller, C.G.; Hawke, D.H.; Tso, J.*; et al.* **(1995)**. Automated C-terminal protein sequence analysis using the Hewlett-Packard G1009A C-terminal protein sequence systems. ed. Crabb, J. *Techniques in Protein Chemistry VI.* San Diego CA, Academic Press**:** 219-27.

Mitchell, P. (**2001**). Microfluidics--downsizing large-scale biology. *Nat Biotechnol* **19**(8): 717-21.

Miyazaki, K.; Tsugita, A. (**2004**). C-terminal sequencing method for peptides and proteins by the reaction with a vapor of perfluoric acid in acetic anhydride. *Proteomics* **4**(1): 11-9.

Morrison, J.R.; Fidge, N.H.; Grego, B. (**1990**). Studies on the formation, separation, and characterization of cyanogen bromide fragments of human AI apolipoprotein. *Anal Biochem* **186**(1): 145-52.

Murphy, C.M.; Fenselau, C. (**1995**). Recognition of the carboxy-terminal peptide in cyanogen bromide digests of proteins. *Anal chem* **67**(9): 1644-45

Mutlu, A.; Pfeil, J.E.; Gal, S. (**1998**). A probarley lectin processing enzyme purified from Arabidopsis thaliana seeds. *Phytochemistry* **47**(8): 1453-9.

Oda, Y.; Nagasu, T.; Chait, B.T. (**2001**). Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat Biotechnol* **19**(4): 379-82.

Oleschuk, R.D.; Harrison, D.J. (**2000**). Analytical microdevices for mass spectrometry. *Trends Anal Chem* **19**(6): 379-88.

Pal, D.; Chakrabarti, P. (**2000**). Terminal residues in protein chains: residue preference, conformation, and interaction. *Biopolymers* **53**(6): 467-75.

Park, H.; Kusakabe, I.; Sakakibara, Y.*; et al.* (**2001**). Autoproteolytic processing of aspartic proteinase from sunflower seeds. *Biosci Biotechnol Biochem* **65**(3): 702-5.

Patrie, S.M.; Ferguson, J.T.; Robinson, D.E.*; et al.* (**2005**). Top down mass spectrometry of <60 kDa proteins from methanosarcina acetivorans using Q-FTMS with automated octopole collisionally activated dissociation (OCAD). *Mol Cell Proteomics*.

Patterson, D.H.; Tarr, G.E.; Regnier, F.E.*; et al.* (**1995**). C-terminal ladder sequencing via matrix-assisted laser desorption mass spectrometry coupled with carboxypeptidase Y time-dependent and concentration-dependent digestions. *Anal Chem* **67**(21): 3971-8.

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

Quadroni, M.; James, P. (**1999**). Proteomics and automation. *Electrophoresis* **20**(4-5): 664-77.

Raffin-Sanson, M.L.; de Keyzer, Y.; Bertagna, X. (**2003**). Proopiomelanocortin, a polypeptide precursor with multiple functions: from physiology to pathological conditions. *Eur J Endocrinol* **149**(2): 79-90.

Rai, D.K.; Landin, B.; Alvelius, G.*; et al.* (**2002**). Electrospray tandem mass spectrometry of intact beta-chain hemoglobin variants. *Anal Chem* **74**(9): 2097-102.

Ramalho-Santos, M.; Verissimo, P.; Cortes, L.*; et al.* (**1998**). Identification and proteolytic processing of procardosin A. *Eur J Biochem* **255**(1): 133-8.

Rappsilber, J.; Mann, M. (**2002**). What does it mean to identify a protein in proteomics? *Trends Biochem Sci* **27**(2): 74-8.

Rosnack, K.J.; Stroh, J.G. (**1992**). C-terminal sequencing of peptides using electrospray ionization mass spectrometry. *Rapid Commun Mass Spectrom* **6**(11): 637-40.

Runeberg-Roos, P.; Saarma, M. (**1998**). Phytepsin, a barley vacuolar aspartic proteinase, is highly expressed during autolysis of developing tracheary elements and sieve cells. *Plant J* **15**(1): 139-45.

Samyn, B.; Hardeman, K.; Van der Eycken, J.*; et al.* (**2000**). Applicability of the alkylation chemistry for chemical C-terminal protein sequence analysis. *Anal Chem* **72**(7): 1389-99.

Schlack, P.; Kumpf, W. (**1926**). Uber eine neue methode zur ermittlung der konstitution von peptiden. *Physiol Chem* **154**: 125-70.

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

Scott, M.G.; Crimmins, D.L.; McCourt, D.W.*; et al.* (**1988**). A simple in situ cyanogen bromide cleavage method to obtain internal amino acid sequence of proteins electroblotted to polyvinyldifluoride membranes. *Biochem Biophys Res Commun* **155**(3): 1353-9.

Sechi, S.; Chait, B.T. (**2000**). A method to define the carboxyl terminal of proteins. *Anal Chem* **72**(14): 3374-8.

Shimizu, T.; Miyairi, K.; Okuno, T. (**2000**). Determination of glycosylation sites, disulfide bridges, and the C-terminus of Stereum purpureum mature endopolygalacturonase I by electrospray ionization mass spectrometry. *Eur J Biochem* **267**(8): 2380-9.

Shively, J.E.; Hawke, D.; Jones, B.N. (**1982**). Microsequence analysis of peptides and proteins. III. Artifacts and the effects of impurities on analysis. *Anal Biochem* **120**(2): 312-22.

Simoes, I.; Faro, C. (**2004**). Structure and function of plant aspartic proteinases. *Eur J Biochem* **271**(11): 2067-75.

Stennicke, H.R.; Mortensen, U.H.; Breddam, K. (**1996**). Studies on the hydrolytic properties of (serine) carboxypeptidase Y. *Biochemistry* **35**(22): 7131-41.

Suckau, D.; Resemann, A. (**2003**). T3-sequencing: targeted characterization of the N- and C-termini of undigested proteins by mass spectrometry. *Anal Chem* **75**(21): 5817-24.

Takamoto, K.; Kamo, M.; Kubota, K.*; et al.* (**1995**). Carboxy-terminal degradation of peptides using perfluoroacyl anhydrides. A C-terminal sequencing method. *Eur J Biochem* **228**(2): 362-72.

Tanaka, S. (**2003**). Comparative aspects of intracellular proteolytic processing of peptide hormone precursors: studies of proopiomelanocortin processing. *Zoolog Sci* **20**(10): 1183-98.

Thiede, B.; Wittmann-Liebold, B.; Bienert, M.*; et al.* (**1995**). MALDI-MS for C-terminal sequence determination of peptides and proteins degraded by carboxypeptidase Y and P. *FEBS Lett* **357**(1): 65-9.

Tsugita, A.; Takamoto, K.; Kamo, M.*; et al.* (**1992**). C-terminal sequencing of protein. A novel partial acid hydrolysis and analysis by mass spectrometry. *Eur J Biochem* **206**(3): 691-6.

Tsugita, A.; Kamo, M.; Miyazaki, K.*; et al.* (**1998**). Additional possible tools for identification of proteins on one- or two-dimensional electrophoresis. *Electrophoresis* **19**(6): 928-38.

van Strien, F.J.; Jespersen, S.; van der Greef, J.*; et al.* (**1996**). Identification of POMC processing products in single melanotrope cells by matrix-assisted laser desorption/ionization mass spectrometry. *FEBS Lett* **379**(2): 165-70.

Vanrobaeys, F.; Devreese, B.; Lecocq, E.*; et al.* (**2003**). Proteomics of the dissimilatory iron-reducing bacterium Shewanella oneidensis MR-1, using a matrix-assisted laser desorption/ionization-tandem-time of flight mass spectrometer. *Proteomics* **3**(11): 2249-57.

Ward, C.W. **(1986)**. Carboxyl terminal sequence analysis. ed. Darbre, A. *Practical Protein Chemistry, A Handbook.* New York, John Wiley & Sons Ltd**:** 491-525.

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**(3): 242-7.

White, P.C.; Cordeiro, M.C.; Arnold, D.*; et al.* (**1999**). Processing, activity, and inhibition of recombinant cyprosin, an aspartic proteinase from cardoon (Cynara cardunculus). *J Biol Chem* **274**(24): 16685-93.

Wilkins, M.R.; Gasteiger, E.; Tonella, L.*; et al.* (**1998**). Protein identification with N and C-terminal sequence tags in proteome projects. *J Mol Biol* **278**(3): 599-608.

Zambelli, B.; Stola, M.; Musiani, F.*; et al.* (**2005**). UreG, a chaperone in the urease assembly process, is an intrinsically unstructured GTPase that specifically binds Zn2+. *J Biol Chem* **280**(6): 4684-95.

Zhou, H.; Watts, J.D.; Aebersold, R. (**2001**). A systematic approach to the analysis of protein phosphorylation. *Nat Biotechnol* **19**(4): 375-8.

Zhou, X.W.; Blackman, M.J.; Howell, S.A.*; et al.* (**2004**). Proteomic analysis of cleavage events reveals a dynamic two-step mechanism for proteolysis of a key parasite adhesive complex. *Mol Cell Proteomics* **3**(6): 565-76.

PART 3

DE NOVO SEQUENCE ANALYSIS

3.1. Introduction

Gel based protein identification protocols in proteomics primarily rely on the peptide mass fingerprinting (PMF) technique in which a protein is digested with an endoprotease of known cleavage specificity. The masses of the resulting peptides are measured by mass spectrometry and matched to peptide masses that have been generated theoretically from proteins in databases. Notwithstanding the fact that the PMF approach is useful for identifying proteins in simple mixtures, the identification of proteins in more complex mixtures often requires partial peptide sequence data obtained by tandem mass spectrometry (MS/MS). In bottom-up approaches, on the contrary, uninterpreted fragmentation spectra of selected peptides are submitted in database searches. Sophisticated algorithms for MS/MS-based protein identification such as SEQUEST (Eng *et al*, 1994) and Mascot (Perkins *et al*, 1999) were developed to match fragmentation spectra with sequences in databases. First, these algorithms generate lists of linear peptide sequences that are isobaric, within specified limits, to the selected precursor. The peptide sequences on these lists are then theoretically fragmented and the in-silico generated spectra matched to the submitted spectra.

While these database search algorithms have proved very useful, their performance is directly related to the quality of the product ion spectra. In many cases, the algorithms will yield scores that are below the accepted thresholds, requiring the data to be manually validated. Because algorithms for MS- and MS/MS-based protein identification require the meticulous matching of experimentally determined peptide masses with peptide masses generated from database entries, their applicability is limited to the identification of proteins within databases. For proteins not contained within sequence databases, it is necessary to determine partial or complete amino acid sequences using either manual or automated *de novo* peptide sequence analysis methods.

ESI MS was used for *de novo* sequencing under various conditions, which include collision-induced dissociation (CID) in the triple quadrupole mass analyzer (Wilm *et al*, 1996) and the QqTOF mass analyzer (Shevchenko *et al*, 1997), resonance excitation in the quadrupole ion trap mass analyzer (Arnott *et al*, 1998), and electron capture dissociation in the Fourier transform ion cyclotron resonance mass analyzer (Horn *et al*, 2000). Regardless of the technique employed, interpretation of MS/MS spectra often requires manual interpretation, which remains prohibitively challenging because of the variation in favored ion fragmentation sites, the chemical nature of amino acid side chains and their relative order in the peptide backbone. MALDI *de novo* sequencing has been carried out using CID in QqTOF instruments (Shevchenko *et al*, 2001; Wattenberg *et al*, 2002) and by resonance excitation of singly charged peptides in a homebuilt MALDI-quadrupole-ion trap mass spectrometer (Zhang *et al*, 2003). In the past, limited sequence information was also obtained through the use of post-source decay (PSD) MALDI-TOF (Kaufmann *et al*, 1996; Gevaert *et al*, 1997). A drawback of peptide sequencing by MALDI is the relatively complex and labor intensive interpretation of fragment ion spectra, mainly due to the many types of fragment ions that arise during fragmentation.

Here, we will briefly discuss the techniques used for *de novo* sequence analysis. For a good understanding and interpretation of MS/MS spectra, knowledge of the mechanisms that result in fragmentation of peptides is required (Part 3.1.1.1). In Part 3.1.1.2, methods to simplify fragmentation spectra thereby allowing easier sequence interpretation, are discussed. Improvements on one such method, N-terminal sulfonation of peptides, are presented. Initially, a case study was performed in which the occurrence of the preferential fragmentation pathways, observed for underivatized peptides, was confirmed for N-terminally

sulfonated peptides (Part 3.1.2). In Part 3.1.3 cross-species protein identification strategies are discussed.

3.1.1. *De novo* sequence analysis

Up to 15 years ago, automated Edman degradation was the preferred method to determine the primary structure of proteins/peptides (Edman *et al*, 1967), but the development of improved mass spectrometric techniques subsequently provided faster and more sensitive tools. However, most algorithms used for spectral interpretation are unable to determine the sequence of a peptide if the corresponding protein is not available in databases. Therefore, different approaches were developed to allow *de novo* sequence analysis from mass spectra. These approaches use either MS spectra of selectively truncated peptides, ladder sequence analysis, or fragmentation spectra. A limitation to *de novo* sequence determination with mass spectrometry, irrespective of the approach used, is that isobaric (Gln/Lys and Phe/Met_{ox}) and isomeric isobaric (Leu/Ile) amino acids are generally not uniquely discerned. Nevertheless, ladder sequence analysis (Wang *et al*, 1996) and, in general, mass spectrometric approaches (Cantin *et al*, 2004) can result in the unambiguous identification of PTMs because of the unique mass shifts they convey.

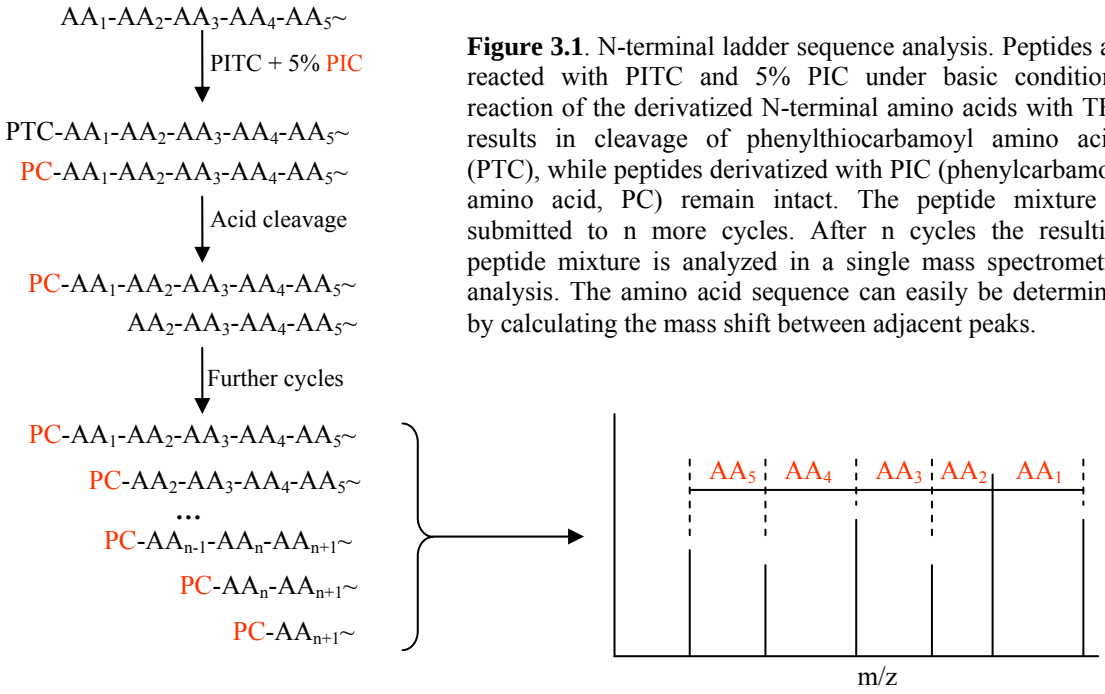


Figure 3.1. N-terminal ladder sequence analysis. Peptides are reacted with PITC and 5% PIC under basic conditions, reaction of the derivatized N-terminal amino acids with TFA results in cleavage of phenylthiocarbamoyl amino acids (PTC), while peptides derivatized with PIC (phenylcarbamoyl amino acid, PC) remain intact. The peptide mixture is submitted to n more cycles. After n cycles the resulting peptide mixture is analyzed in a single mass spectrometric analysis. The amino acid sequence can easily be determined by calculating the mass shift between adjacent peaks.

Ladder sequence analysis involves the sequential truncation of peptides from the C- or the N-terminus. The resulting peptide mixture is analyzed by mass spectrometry, resulting in spectra that display a ladder-like succession of ions corresponding to the sequentially truncated peptide. Sequence is determined by calculating the mass differences between successive peaks. N-terminal ladder sequence approaches have always focused on Edman-like chemistries; for C-terminal ladder sequence analysis both enzymatic and chemical approaches have been used (Bergman, 2000). Examples of C-terminal enzymatic ladder sequence analysis and a more elaborate discussion are given in Part 2. N-terminal ladder sequence analysis was pioneered by Chait & Wang, and an overview of their approach is presented in Figure 3.1 (Chait *et al*, 1993).

Because doubly charged peptide ions tend to fragment more equally across a given sequence than do singly charged ion species (Cramer *et al*, 2001; Tabb *et al*, 2003), a large proportion of *de novo* sequencing has been performed on doubly charged ions, which are readily produced from tryptic peptides by ESI. In contrast, MALDI produces predominantly singly charged ions from peptides. These singly charged ions often yield relatively low quality CID mass spectra which tend to be dominated by a few preferred fragmentation pathways (Qin *et al*, 1995). Nevertheless, fragmentation of singly charged peptide ions became routine with the introduction of hybrid MALDI Q-TOF (Loboda *et al*, 2000; Shevchenko *et al*, 2000), MALDI-TOF/TOF (Medzihradszky *et al*, 2000) and MALDI LIFT-TOF/TOF (Suckau *et al*, 2003a; Suckau *et al*, 2003b) instruments. The comparison of fragmentation spectra of doubly and singly charged precursor ions reveals striking differences: peptide fragment ions obtained from doubly charged precursors are mainly y-type ions and some b-ions, whereas peptide fragment ions produced from singly charged ions are a mixture of y-, b- and a-ions, accompanied by ions resulting from the neutral loss of ammonia or water. Furthermore for MALDI generated ions, the ratio and intensity of these fragment ions is strongly sequence-dependent, hampering *de novo* sequence analysis.

Apart from the occurrence of different incomplete ion series in fragmentation spectra, the main problem associated with MS/MS-based *de novo* sequence determination is one of directionality. It is a priori impossible to determine which fragment ions belong to the b-ion series and which ones to the y-ion series. The difficulty of *de novo* sequence analysis, and in particular of identifying the respective ion series, has led to several ingenious methods to solve this problem. For instance, fragment ions belonging to a y-ion series can be identified by performing proteolysis with trypsin in a buffer containing 50% $H_2{}^{18}O$/50% $H_2{}^{16}O$ (v/v). During hydrolysis with proteases, a molecule of water is added to each amide bond that is hydrolyzed. If both differentially isotopically labeled parent ions are selected, only fragment ions that contain the intact carboxyl-terminus will appear as doublets separated by 2 thomson units (Gaskell *et al*, 1988; Gevaert *et al*, 1997). Similar results were obtained by esterification of all carboxyl groups, including the C-terminus (Hunt *et al*, 1986). However, the latter method requires that MS/MS spectra are acquired before and after esterification. Other chemistries that result in facilitated *de novo* sequence analysis by differentiation of N- and C-terminal fragments have been proposed, involving either the introduction of a label during cell culturing (Gu *et al*, 2002; Gu *et al*, 2003; Shui *et al*, 2005) or derivatization of peptides after proteolytic digestion (Munchbach *et al*, 2000; Brancia *et al*, 2004; Beardsley *et al*, 2005).

The development of software that allows *de novo* sequence determination based on fragmentation spectra of peptides is another possibility to improve MS/MS approaches. Because it would eliminate the need for *de novo* sequence analysis, creation of random databases, containing all possible peptide sequences, is conceptually the simplest solution. However, attempts to do that failed, due to the tremendous number of possible amino acid arrangements (Lu *et al*, 2004). Lutefisk (Taylor *et al*, 2001), Pepnovo (Frank *et al*, 2005) and NovoHMM (Fischer *et al*, 2005) are just some of the algorithms that were recently developed. Fischer *et al* used the different algorithms to analyze the same data set and compared the results. A general observation from this comparison, neglecting small differences in performance of the individual algorithms, is that short sequences of 3 to 4 amino acids are correctly determined for more than 70% of the spectra. However, longer correct sequences were infrequently determined. Although sequences of on average 10.3 amino acids were called, none of the 5 algorithms was able to determine a correct sequence of 10 or more amino acids for more then 20% of the spectra (Fischer *et al*, 2005). Similar results were obtained using a algorithm for sequence optimization, as opposed to *de novo* sequence determination

(Heredia-Langner *et al*, 2004). In general, algorithms for automated *de novo* sequence determinations result in a number of similar, redundant sequences with marginally different scores. This is illustrated in Table 3.1, wherein the *de novo* determined sequences for two peptides from cytochrome c are depicted, fragmented with the 4700 MALDI-TOF/TOF mass spectrometer. The sequences in Table 3.1 were determined with the DeNovoExplorer interface that is integrated in the software package of the instrument.

Table 3.1. *De novo* determined sequences using DeNovoExplorer

	1168.56 Da (TGPNLHGLFGR)	Score	1584.84 Da (KTGQAPGFSYTDANK)	Score
1	TGPNLHGLMGR	83.76	KTGTCAGLMYAGPNK	48.99
2	NPGTLHGLMGR	81.72	KTGTCAGLMYPGANK	48.20
3	PGGGTLHGLMGR	81.02	KTGCTAGLMYAGPNK	47.94
4	GPGGTLHGLMGR	81.02	KTGCGDGLMYPGANK	47.80
5	TGNPLHGLMGR	80.03	KTGCGDGLMYAGPNK	47.76
6	TGPNIHGLFGR	79.77	KTTGCAGLMYAGPNK	47.46
7	TGPNLHVAMGR	78.84	KTGCTAGLMYPGANK	47.39
8	GPNTLHGLMGR	78.46	KTASCAGLMYAGPNK	47.07
9	TGPNLHAVMGR	77.83	KTGTCAGYPYAGPNK	46.42
10	NPGTIHGLFGR	77.83	KTGTCAGLMYAGAGPN	46.39

De novo determined sequences using DeNovoExplorer: trypsin was specified as enzyme, and the oxidation of methionine and carboxymethylamidation of cysteine as allowed variable modifications. A mass error of 0.2 Da was tolerated. Residues indicated in red are correct throughout the 10 best scoring sequences. The difference in score between the highest and the lowest scoring sequence is 7% for the peptide at 1168.56 Da (highest hit: 11/11 amino acids correctly assigned versus 7/11 for the lowest) and 5.5% for the peptide at 1584.84 Da (highest scoring sequence: 6/15 amino acids correct, lowest: 3/15). The spectra were also submitted in a Mascot-search (using identical settings) against the entire NCBI-database and resulted in a correct and significant identification.

To improve the performance of algorithms for *de novo* sequence determinations, the rules that govern fragmentation of peptides should be encrypted in the software. This would allow the use of peak intensity as a parameter for assigning a sequence to a fragmentation spectrum. The benefits of using a peak-intensity-matrix in a database search algorithm, i.e. VEMS (Matthiesen *et al*, 2005), were recently presented (Hjerno *et al*, 2005). In this poster only improvements during database-dependent sequence determination are reported. However, similar improvements can be expected when this matrix is integrated in algorithms for *de novo* sequence analysis.

3.1.1.1. Fragmentation pathways

Currently, the most common type of fragmentation used in protein research is low-energy CID. In this technique peptide ions of a particular mass-to-charge ratio are selected and excitated. Excitation, either by collision with neutral gas atoms or by electrostatics, adds energy to the peptide ion, energy that is dissipated by fragmentation. In the 'mobile proton model', cleavage of peptide bonds is thought to be mediated by the mobility of the ionizing proton (Dongré *et al*, 1996). Although this model allows to describe the probability of cleavage of particular bonds, it only offers a mechanistic model; the intensity of a particular fragment ion in fragmentation spectra is not explained. Besides the probability of cleavage, the presence of a fragment ion in MS/MS spectra depends on the proton affinity of the fragments, and on the energetics and the kinetics of the fragmentation. These factors are considered in the 'pathways in competition model' (PIC-model), a more comprehensive model that is currently under development (Paizs *et al*, 2005).

Although it only describes predissociation events, the 'mobile proton model' is now generally accepted. The model postulates that, upon excitation, the proton(s) added to a peptide will migrate to various protonation sites prior to fragmentation, provided they are not sequestered by a basic amino acid side chain (Wysocki *et al*, 2000). A concept that was pioneered with the introduction of the 'heterogeneous population model' in 1992 (Burlet *et al*, 1992). Experimentally, the mobile proton model was verified by deuterium labeling techniques (Johnson *et al*, 1995; Harrison *et al*, 1997) which indicate strong H/D mixing prior to collisionally activated dissociation. The recently described randomization of deuterium labels among all N- and O-linked hydrogens during both low- and high-energy fragmentations confirms the mobility of protons during MS/MS analysis (Jorgensen *et al*, 2005a; Jorgensen *et al*, 2005b).

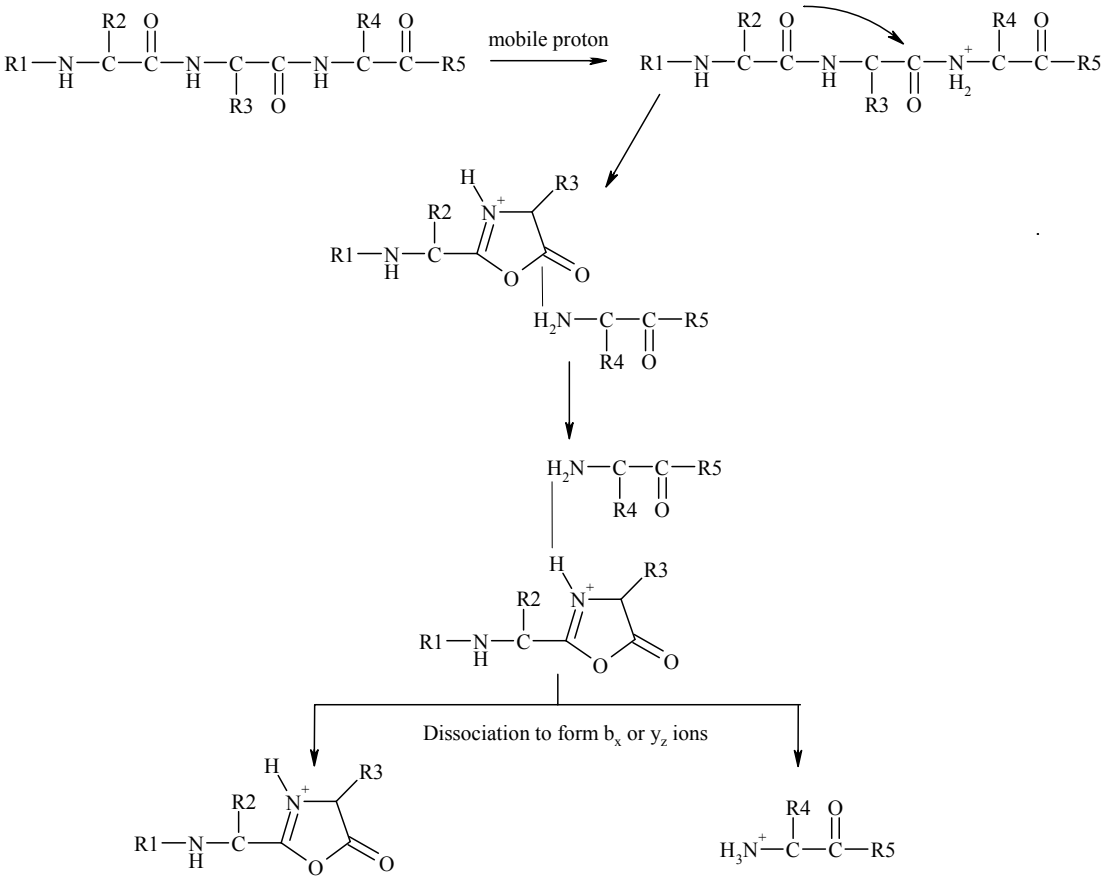


Figure 3.2. Fragmentation of peptide bonds according to the mobile proton model. Nucleophilic attack of the carbonyl oxygen of the N-terminal neighbor peptide bond on the carbon center of the protonated amide bond results in the formation of an oxazolone and dissociation of the peptide bond. Dissociation of the proton-bound complex that is subsequently formed results in formation of b- or y-ions depending on the proton affinity of the fragments.

The factors determining the mobility of the proton were studied, and a higher onset for fragmentation was noticed for peptides containing basic residues (Jones *et al*, 1994). As the peptide contains a more basic group (AAAAA<PAAAA<KAAAA<RAAAA, mimicking the increased gas-phase basicity) the energy required to 'mobilize' the proton increases (Dongré *et al*, 1996). The main characteristics of the mobile proton model are; firstly that fragmentation is initiated by protonation of the corresponding peptide bond. Secondly, although protonation of the amide oxygen is thermodynamically favored, this will increase the peptide bond strength. Therefore, protonation of amide nitrogen is thought to initiate fragmentation, resulting in a decrease of the peptide bond strength and rendering the carbonyl

carbon a possible target for nucleophilic attack (Somogyi *et al*, 1994). Thirdly, the extent of fragmentation and the energy required to induce it depends on the presence of basic groups in the peptide.

According to the mobile proton model, fragmentation is thus initiated by protonation of peptide bond, and the weakened peptide bond is more vulnerable to cleavage. Furthermore, the carbon atom of the amide bond is a likely target for nucleophilic attack of close-by electron-rich groups. The mechanism that was proposed for the subsequent cleavage reaction is depicted in Figure 3.2. Experimental and theoretical evidence that supports this mechanism is accumulating. For instance, the proposed structure of b-ions, protonated oxazolones, was confirmed by ab-initio calculations and secondary fragmentation experiments (Yalcin *et al*, 1995; Nold *et al*, 1997; Paizs *et al*, 2003; Chen *et al*, 2005).

Although the mobile proton model allows to predict whether or not a specific bond is fragmented, it does not provide clues on the intensity of the resulting fragments in MS/MS spectra. The PIC-model integrates mechanistic with kinetic and energetic descriptions of the different fragmentation pathways. Although in its infancy, the use of this model allowed the fairly accurate quantitative prediction of fragmentation spectra for oligoalanines (Harrison *et al*, 2004), by considering the proton affinities of the different fragmentation products (Paizs *et al*, 2004). Recently, statistical evaluations of the intensity of fragment ions in large MS/MS data sets reveal tendencies that will allow the refinement of the PIC-model (Huang *et al*, 2002; Kapp *et al*, 2003; Tabb *et al*, 2003; Tabb *et al*, 2004).

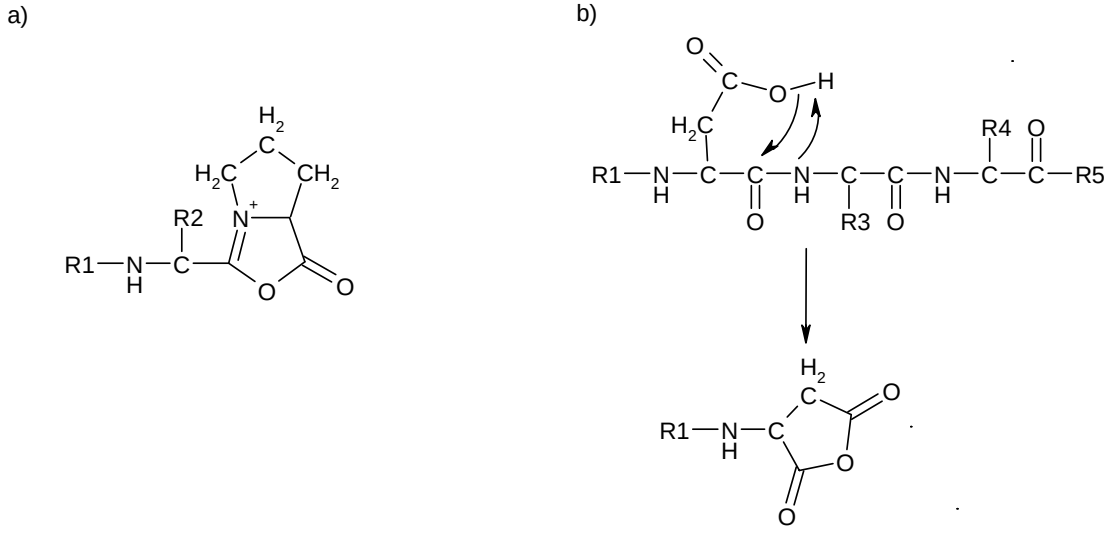


Figure 3.3. a) The highly strained bicyclic structure that would be formed during fragmentation of peptide bonds C-terminal to Pro according to the mechanism in Figure 3.2. b) The concerted mechanism for the enhanced cleavage of peptide bonds C-terminal to Asp.

Individual amino acids residues influence which of the two adjacent amide bonds (N- or C-terminal) break in the fragmentation process (Breci *et al*, 2003). The structure of Pro, for example, prevents the cleavage of the peptide bond C-terminal to this residue because cleavage of this bond would require the formation of a highly strained bicyclic structure (Figure 3.3a) (Vaisar *et al*, 1996). The biases of amino acid residues to result in fragment ions corresponding to N- or C-terminal peptide bond fragmentation were revealed in a statistical analysis of a large MS/MS data set (Tabb *et al*, 2003). In the y-ion series, a strong N-terminal bias was established for Pro and Gly. Although this analysis was performed on MS/MS

spectra of doubly charged precursors, the gaps C-terminal to Pro and Gly observed in fragmentation spectra of sulfonated peptides corroborate these preferences.

The most dominant fragment ion peak in MS/MS spectra of singly charged tryptic peptides often corresponds to the cleavage of the peptide bond C-terminal to aspartic acid. The preferred fragmentation of this peptide bond, and to a lesser degree of the Glu-Xxx bond (Gu *et al*, 2000), is thought to be the result of a charge-remote fragmentation pathway (Tsaprailis *et al*, 2000; Wysocki *et al*, 2000). Several mechanisms have been proposed to explain this specific cleavage, all starting with the transfer of the acidic proton from the aspartic acid side chain to the amide nitrogen of the peptide bond C-terminal to the acidic amino acid. Subsequent nucleophilic attack of the side chain carbonyl on the peptide bond results in peptide bond cleavage and the formation of a b-ion with a cyclic anhydride structure (Yu *et al*, 1993). Only a concerted mechanism, as depicted in Figure 3.3b, can explain the absolute preferential fragmentation of the Asp-Xxx peptide bond (Tsaprailis *et al*, 1999; Paizs *et al*, 2002).

Pathways, similar to those for peptide backbone fragmentation, have been proposed for neutral losses. Although neutral losses generally provide little sequence defining information, some notable exceptions are known. The neutral loss of 64 Da ($-CH_3SOH$) from the side chain of oxidized methionine allows to unambiguously distinguish this modified amino acid derivative from the isobaric amino acid phenylalanine (Lagerwerf *et al*, 1996). The neutral loss of H_2O and NH_3 is frequently observed; loss of water can occur at carboxyl groups (C-terminal, Asp and Glu) and from Ser and Thr side chains (Ballard *et al*, 1993). Loss of ammonia on the other hand occurs at the side chains of Asn, Gln, Lys and Arg, but not from the N-terminus. During fragmentation of N-terminally sulfonated peptides containing an internal basic residue, we always observed the neutral loss of NH_3 from the side chains of internal basic residues homoarginine (after guanidination of Lys, see Figure 3.6) and arginine (Martin *et al*, 2005). Up to the place where the internal basic residue was fragmented, the peaks in the (y-17)-ion series were often more pronounced than peaks in the corresponding y-ions (e.g. Part 3.2, Figure 3.16). This knowledge allowed us to resolve some ambiguous sequence calls. Because peaks corresponding to the fragmentation of the peptide bond C-terminal to Gly are often not observed (Tabb *et al*, 2003), one sequence that can lead to erroneous sequence calls is the dipeptide Gly-Val, isobaric with Arg (156.1 Da) (Nielsen *et al*, 2005; Savitski *et al*, 2005). However, when no neutral loss of 17 Da was observed in the y-ion series, the sequence Gly-Val was called for the 156 Da mass differences between apparently consecutive y-ions.

3.1.1.2. Charge derivatization

Another approach to facilitate MS/MS-based *de novo* sequence determination is to simplify MS/MS spectra. Methods that allow to distinguish N- from C-terminal fragments are helpful for sequence determination using ESI-MS of multiple charged peptides that fragment at low internal energies, generally below 300 eV. However, because higher fragmentation energies are required to fragment singly charged peptides (Wysocki *et al*, 2000), these methods are not sufficient to allow easy *de novo* sequence determination of singly charged peptides. High-energy peptide fragmentation typically results in complex fragmentation spectra with different incomplete fragment ion series, often of low intensity. The influence of the position of the charge, the most basic functionality, on the ion types that are formed during fragmentation of singly charged peptides was described in 1988 (Johnson *et al*, 1988). When a basic group is positioned at the N-terminus of the peptide, fragmentation

predominantly results in a_n, b_n and d_n ions, whereas a C-terminal basic group primarily results in w_n, y_n and z_n ions. Mixtures of N- and C-terminal ion types are present when the most basic group is situated in the centre of the peptide. Apart from the position of the basic group, the type of fragment ions that are formed is also influenced by the basicity (proton affinity) of groups in the peptide. This influence is schematically represented in Figure 3.4. When groups of higher basicity are present in a peptide, the charge is more retained at that specific residue. Following the mobile proton model for mass spectrometric fragmentation, this increased retention of charge has an impact on the type of ions that are formed.

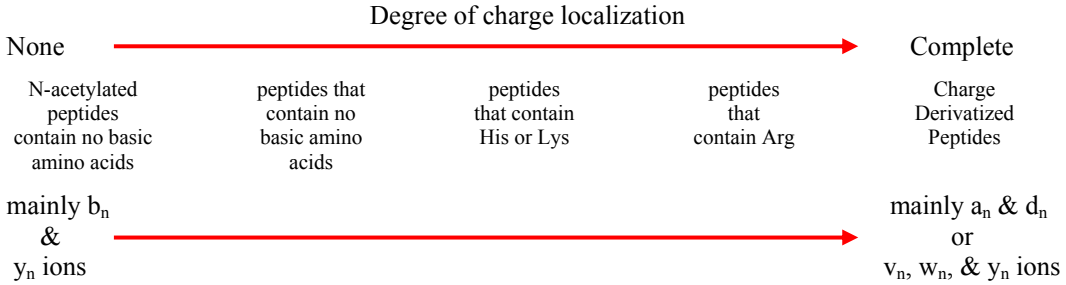


Figure 3.4. Dependence of the charge localization on the most basic group in peptides. Dependent on the degree of charge fixation, more energy is required to induce charge-driven fragmentation, resulting in the formation of different types of fragment ions during charge-remote fragmentation (Johnson *et al*, 1988).

Therefore, it was postulated that derivatization of one of the termini of peptides with a fixed charge directs fragmentation, thereby resulting in the simplification of fragmentation spectra (Stults *et al*, 1993). Two approaches for charge derivatization have been developed. The first is derivatization of peptides with a fixed charge that limits the mobility of the ionizing proton, favoring charge-remote fragmentation (Part 3.1.1.2.1)(Roth *et al*, 1998). In the second approach (Part 3.1.1.2.2), the positive charge of the C-terminal basic amino acid of tryptic peptides is neutralized by N-terminal derivatization with a negatively charged group. Consequently, the ionizing proton of the protonated peptide is freely mobile and promotes charge-driven fragmentation (Figure 3.5) (Keough *et al*, 1999).

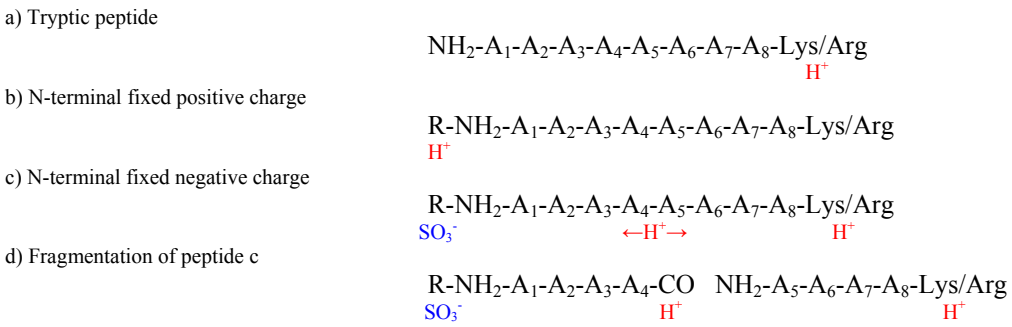


Figure 3.5. The different types of derivatization. a) Singly charged tryptic peptide protonated at the most basic site. b) N-terminal positive charge derivatization fixates the proton at the N-terminus, inducing charge-remote fragmentations. c) Derivatization of the N-terminus of a tryptic peptide with a negative charge allows the ionizing proton to move freely along the peptide backbone and induces charge-driven fragmentations. d) Products after gas phase fragmentation of the peptide depicted in (c), the N-terminal b-type fragment is neutral and not observed in the MS/MS spectra.

3.1.1.2.1. Fixed charge

As can be deduced from Figure 3.4, charge fixation has an impact on the types of fragments that are formed during tandem mass spectrometric experiments. The different

reagents used for the introduction of fixed charges and their effect on the fragmentation have been reviewed in 1998 (Roth *et al*, 1998). Most of the techniques described in this review involve derivatization of the N-terminus of a peptide with a cationic group, resulting in the dominant presence of a-type ions. An important reason to favor cationic derivatization, N-terminal attachment of a quaternary ammonium (Kidwell *et al*, 1984; Stults *et al*, 1993) or phosphonium group (Wagner *et al*, 1991), was that derivatization should enhance the sensitivity of MS and MS/MS analysis, using relatively insensitive Fast Atom Bombardment (FAB) ionization. With the better detection limits offered by MALDI-ionization and the development of post source decay (PSD) (Spengler *et al*, 1992) this incentive disappeared.

Methods have also been developed to derivatize the C-terminal carboxyl group. While the difference in pKa between the primary amino group and the ε-amino group of Lys is sufficient to allow specific derivatization, no such pKa difference exists for carboxyl groups. Therefore, reactions that convert the C-terminal carboxyl group to a reactive electrophile, also used for the activation in C-terminal chemical sequence analysis (Part 2.1.), have been proposed to overcome this lack in specificity. Derivatization of peptides with fixed negative charges was performed using amino naphthalene sulfonic acid. As for other proposed derivatization reactions, the utility of this approach was hampered by side reactions and the limited benefit that resulted from this derivatization (Lindh *et al*, 2000).

In general, positive fixed charge derivatization results in increased mass spectrometric sensitivity and a limited reduction of the complexity of MS/MS spectra. Nevertheless, no fixed charge derivative has ever been described that complies to all criteria for an ideal derivatization strategy (Roth *et al*, 1998). Three different peptide derivatives for fixed charge derivatization approach these criteria. The dimethylalkylammonium acetyl (DMAA) derivative (Stults *et al*, 1993), the C5Q derivative (Bartlet-Jones *et al*, 1994) and the [tris(trimethoxyphenyl)phosphonium] acetate (TMPP) derivative (Huang *et al*, 1997) use highly specific derivatization chemistries to attach a cationic group to the N-terminus of peptides with high yields. Of these three chemistries, TMPP-derivatization is most frequently used (Liao *et al*, 1997; Strahler *et al*, 1997; Shen *et al*, 1999; Sadagopan *et al*, 2000; Czeszak *et al*, 2004)

3.1.1.2.2. Formation of singly charged peptides with a mobile proton

Gaskell and coworkers (Burlet *et al*, 1992; Burlet *et al*, 1995; Cox *et al*, 1996) demonstrated that the oxidation of cysteine to cysteic acid (Figure 3.6a) in peptides containing a C-terminal Arg increases the yield of y-type fragment ions observed by CID of protonated molecules. This idea was the starting point for the research group of Keough to develop a general procedure for high-sensitivity tryptic peptide sequencing using PSD-MALDI (Keough *et al*, 1999).

By performing N-terminal sulfonation, the charge of the strong acidic group counterbalances the C-terminal positive charge. In the sulfonated peptides, the ionizing proton will be more or less free to randomly ionize the backbone amide groups, as the most basic residue is already protonated (Figure 3.5c). In the resulting PSD spectra, increased fragmentation by direct cleavage of protonated amide bonds has been observed. The major products of these fragmentation reactions have structures as depicted in Figure 3.5d. Fragments containing the N terminus, derivatized b-ions with a protonated oxazolone, will be neutral and suppressed in the positive ion mode. Only C-terminal y-ion fragments will be enhanced because they contain the protonated basic C-terminal amino acid of tryptic peptides.

Only y-ions will be observed in the resulting fragmentation spectra, which allows facile sequence determination by calculating the mass difference between consecutive peaks.

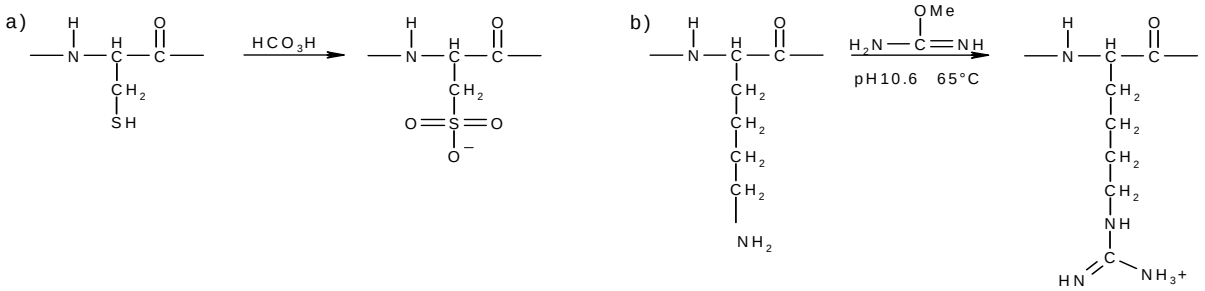


Figure 3.6. Structure of two side chain modifications. a) Complete oxidation of cysteine to cysteic acid can be performed with different strong oxidants; oxidation with performic acid is most often used. b) Conversion of the lysine side chain into the more basic homoarginine (Beardsley *et al*, 2002).

In their first article, Keough *et al* investigated the use of different acidic groups. Derivatization with carboxyl acid containing groups resulted in spectra comparable to those of underivatized peptides. Because of their relatively high pKa (ranging from 1.8 to 4.2), these derivatives were protonated during MALDI analysis. Only derivatization with strong acidic groups, such as sulfonic acids, resulted in the formation of singly charged peptides with a mobile proton and in the suppression of N-terminal fragments ions.

The different reagents used for N-terminal sulfonation are also reactive towards other amino groups and it was observed that disulfonate derivatives are undesirable; they exhibit poor sensitivity in positive-ion mode and poor fragmentation behavior under negative-ion conditions. Therefore, the original concept was only applicable on arginine containing tryptic peptides. In the guanidination reaction, the lysine side chain is converted to a guanidino group (Figure 3.6b), having a higher pKa, thereby preventing reaction with the reagents used for sulfonation (Keough *et al*, 2000b). Furthermore, guanidination of lysine allows the differentiation of this residue, with a residual mass of 170 Da as homoarginine, from the isobaric amino acid glutamine. Guanidination is a frequently used technique in protein chemistry (Kimmel, 1967); amongst others it is used to stabilize protein structures by providing stronger salt bridges (Cupo *et al*, 1980) and as a tool to enhance sequence specific fragmentation (Bunk *et al*, 1993). Because MALDI-PMF spectra are generally dominated by peptides that contain a C-terminal arginine (Krause *et al*, 1999), guanidination has been applied in a number of methods to increase the mass spectral signal intensities from lysine containing peptides (Beardsley *et al*, 2000; Brancia *et al*, 2000; Hale *et al*, 2000). Brancia *et al* combined the benefits of increased sensitivity of MALDI-PMF of guanidinated lysine containing peptides with N-terminal derivatization of peptides with phenylisothiocyanate (Brancia *et al*, 2001). The latter modification induces selective dissociation of N-terminal peptide bonds in gas-phase fragmentation experiments, enabling the facile identification of N-terminal residues (Summerfield *et al*, 1999). Because it allows to differentiate Lys from Gln, guanidination was also applied for enzymatic C-terminal sequence analysis (Bonetto *et al*, 1997). In 2002, an optimized protocol for the guanidination of peptides was reported. Reaction times for guanidination were typical from 1 to 16 hours and the reaction often resulted in incomplete modification. Optimization of the reaction temperature and the concentration of *O*-methylisourea resulted in complete conversion of lysine side chains in 5 minutes (Beardsley *et al*, 2002). The only side reaction noted is the guanidination of the N-terminal amino group if the N-terminal residue is Gly (Beardsley *et al*, 2000; Cotter *et al*, 2001).

Initially, Keough *et al* used two different reagents for N-terminal sulfonation: 2-sulfobenzoic acid cyclic anhydride and chlorosulfonylacetyl chloride (Figure 3.7a & b). They reported the use of the 2-sulfobenzoic acid reagent as most preferable because it can be used in aqueous solution. Derivatization with chlorosulfonylacetyl chloride in water resulted in lower derivatization yields; the reagent reacts with water and therefore must be used in dry THF. After derivatization in THF, a chlorosulfonyl peptide is formed which rapidly converts to the wanted sulfonic acid peptide after addition of water. Later on, Keough *et al* preferentially used chlorosulfoacetyl chloride, and the derivatized peptides were analyzed using different types of mass spectrometers (Bauer *et al*, 2000; Keough *et al*, 2000a; Keough *et al*, 2000b; Lacey *et al*, 2000; Keough *et al*, 2001). In 2002 a new reagent was proposed, 3-sulfopropionic acid succinimidyl ester (Figure 3.7c), that allows to perform the entire protocol in aqueous solutions (Keough *et al*, 2002).

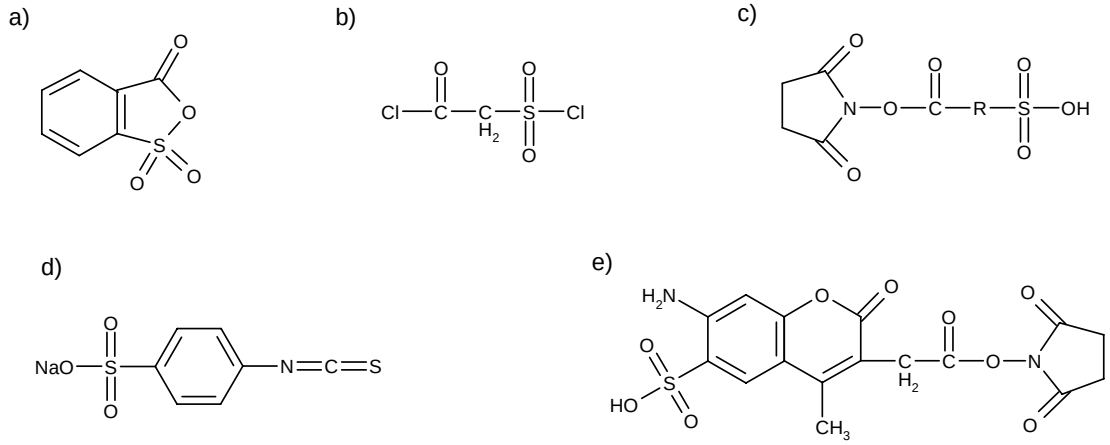


Figure 3.7. Different reagents used for N-terminal sulfonation of peptides; a) 2-sulfobenzoic acid cyclic anhydride, b) chlorosulfonylacetyl chloride, c) 3-sulfopropionic acid succinimidyl ester with R=C_2H_4 (now commercially available in the Ettan CAF-MALDI Sequencing Kit, GE Healthcare), d) 4-sulfophenyl isothiocyanate (SPITC) e) Alexa Fluor® 350 carboxylic acid, succinimidyl ester (Invitrogen).

During guanidination, a large molar excess of *O*-methylisourea semisulfate is used. Therefore, the samples require cleanup prior to MALDI analysis, a step that is typically done on C18 mini-columns. Loading of these columns requires that the sample is dissolved in aqueous solution. Therefore, samples sulfonated in organic solvent have to be dried, inducing possible sample loss, and redissolved in water. Initially, derivatization with water-compatible reagents was slow because of the lower reactivity of the reagents and the low concentration of analytes. Solid-phase sulfonation, after concentration of the peptides on C18 micro-purification tips, resulted in near complete derivatization in seconds and allowed multiplexing (Keough *et al*, 2002; Keough *et al*, 2003). The 3-sulfopropionic acid succinimidyl ester reagent was commercialized by Amersham Bioscience as a kit named 'Chemically Assisted Fragmentation' (CAF). After solid-phase sulfonation, the sample is cleaned up and the peptides either eluted directly on MALDI-probes (Hellman *et al*, 2002; Keough *et al*, 2003) or submitted to subsequent chromatographic separations (Flensburg *et al*, 2005).

Prior to the introduction of the 3-sulfopropionic acid succinimidyl ester in 2002, another water-compatible reagent has been reported. In 2001, Gevaert *et al* used 4-sulfophenyl isothiocyanate (SPITC) (Figure 3.7d) for the derivatization of a synthetic peptide (Gevaert *et al*, 2001). This allowed to almost completely determine the sequence of the derivatized peptide YSFVATAER in PSD spectra. However, because a rather high amount of peptide was required, this derivatization was considered unsatisfactory and only amendable for particular cases. Derivatization with SPITC was used again in 2003, resulting in

apparently quantitative derivatization, judged by the absence of peaks corresponding to underivatized peptides, of the peptides from a tryptic digestion of a protein (Marekov *et al*, 2003). Gevaert *et al* performed the derivatization in aqueous solution, Marekov & Steiner, on the contrary, used a mixture of ethanol, triethylamine and water (80/11/9) (Table 3.2). A similar reaction mixture, in which the poorly volatile triethylamine was replaced with pyridine (Lee *et al*, 2004b), was used for quantitative analysis of sulfonated peptides (Lee *et al*, 2004a). A study, wherein sulfonated peptides were selectively isolated using a fullerene-derivative (C60) also applied this reaction mixture (Lee *et al*, 2006). However, a systematic study of the best reaction conditions, optimizing both the reaction yields and the effects on mass spectrometric analysis, showed that reaction in an aqueous environment is preferable (Wang *et al*, 2004). So far, protection of lysine by guanidination has not been reported in any of the SPITC-applications. However, it has been shown that the use of SPITC results in the modification of lysine side chains (Oehlers *et al*, 2005). The decreased sensitivity of mass spectrometric analysis of SPITC-sulfonated peptides can be reversed by the use of a 2,4,6-trihydroxyacetophenone matrix combined with diammonium citrate (Oehlers *et al*, 2005), a matrix composition that has previously been used to increase the sensitivity of MALDI for phosphorylated peptides (Yang *et al*, 2004).

Table 3.2. Reaction conditions applied to peptide derivatization with SPITC

(Gevaert *et al*, 2001)	50 mM Na_2CO_3	pH 8.5	60 minutes	55°C
(Marekov *et al*, 2003)	Ethanol/Triethylamine/Water (80/11/9)		10 minutes	55°C
(Lee *et al*, 2004a)	pyridine/water/ethanol (1/1/2)	pH 8	60 minutes	50°C
(Wang *et al*, 2004)	20 mM $NaHCO_3$	pH 9.5	30 minutes	55°C

The application of two other reagents for N-terminal sulfonation of peptides was described. A method for N-terminal sequence analysis was developed that uses sulfonation. After reduction and alkylation of cysteine residues, lysine side chains are guanidinated and the protein is derivatized at its N-terminus with biotinylcysteic acid (BCA). After trypsin digestion, the derivatized peptide can be isolated taking advantage of the specific avidin-biotine interaction (Yamaguchi *et al*, 2005). Pashkova *et al* reported that derivatization of peptides with different coumarin containing tags results in an increased intensity of the peaks during MALDI analysis (Pashkova *et al*, 2004). One of the coumarin tags they used is the dye Alexa Fluor 350 (Invitrogen) (Figure 3.7e). The MS intensity of Alexa-tagged peptides was better than that of the corresponding CAF-tagged peptide. In a further study, the MS/MS fragmentation pattern was found to be comparable using both reagents (Pashkova *et al*, 2005).

3.1.2. Preferential fragmentation pathways during *de novo* sequence analysis of N-sulfonated peptides

A case study of *de novo* sequence analysis of N-sulfonated peptides by MALDI TOF/TOF mass spectrometry

Bart Samyn, Griet Debyser, Kjell Sergeant, Bart Devreese, and Jozef Van Beeumen

University of Gent

Department Biochemistry, Physiology and Microbiology

Laboratory of Protein Biochemistry and Protein Engineering

K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

Published in J Am Soc Mass Spectrom (**2004**), 15(12), 1838-52**.**

Introduction

The simplicity and sensitivity of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry have increased its application in recent years. The most common method of 'peptide mass fingerprint' analysis often does not provide robust identification. Additional sequence information, obtained by post-source decay or collision induced dissociation, provides additional constraints for database searches. However, *de novo* sequencing by mass spectrometry is not yet common practice, most likely because of the difficulties associated with the interpretation of high and low energy CID spectra. Success with this type of sequencing requires full sequence coverage and demands better quality spectra than those typically used for database searching.

In this part it is shown that full-length *de novo* sequencing is possible using MALDI TOF/TOF analysis. The interpretation of MS/MS data is facilitated by N-terminal sulfonation after protection of lysine side (Keough *et al*, 1999). Reliable *de novo* sequence analysis has been obtained using sub-picomol quantities of peptides and peptide sequences of up to 16 amino acid residues in length have been determined. The simple, predictable fragmentation pattern allows routine *de novo* interpretation. Characterization of the complete primary structure of a peptide is often hindered due to differences in fragmentation efficiencies and in specific fragmentation patterns for different peptides. These differences are controlled by various structural parameters including the nature of the residues present. The influence of the presence of internal Pro, acidic and basic residues on the TOF/TOF fragmentation pattern will be discussed, both for underivatized and guanidinated/sulfonated peptides.

Experimental

Materials

Horse heart cytochrome c (purity 97%), bovine milk α-casein (purity 85%), synthetic peptides and CHCA were from Sigma (Bornem, Belgium). Endoprotease Lys-C was from Roche (Brussels, Belgium) and trypsin was from Promega (Leiden, The Netherlands). Dry tetrahydrofuran and 2-sulfobenzoic acid cyclic anhydride were from Fluka (Bornem, Belgium). HPLC-grade acetonitrile (ACN) was obtained from BioSolve (Valkenswaard, The Netherlands).

Tryptic digestions of proteins

The peptides used in this study were obtained from the proteolytic digestion of two test proteins: horse heart cytochrome c and bovine milk α-casein. Both proteins were incubated with endoprotease LysC and trypsin in 10 mM NH_4HCO_3, pH 8, for four hours at 37°C, at an E/S ratio of 1/40 (w/w). After digestion, the sample was acidified (1 % TFA/MQ), centrifuged, and separated on an analytical RPLC column (Brownlee, 2.1 x 220 mm, 5 μm, C18) using a SMART system (Pharmacia, Uppsala, Sweden). Individual peptide fractions were collected manually in 500 μl Eppendorf vials and dried in a Speedvac (Savant). All fractions were redissolved in 0.1%TFA/MQ at an estimated concentration of 25 pmol/μl. The individual RPLC fractions were analyzed on the 4700 Proteomics Analyzer (Applied Biosytems) in the positive reflectron mode. Only those fractions containing one or two principal peptide components were judged to be pure and used in further experiments (Table 1).

Guanidination and sulfonation

An *O*-methylisourea stock solution was prepared by dissolving 50 mg of the compound in 51 μl of MQ. A freshly prepared *O*-methylisourea stock solution was used in every reaction. For the guanidination modification 1.5 μl of this stock solution was mixed with 2 μl of the peptide solution (50 pmol), 3 μl MQ, 5.5 μl 7N NH_4OH, and incubated for 10 minutes at 65°C. After guanidination, the peptides were desalted on a Prosorb device (PVDF) with two subsequent washes of 100 μl MQ, and extracted with 10 μl 50 % ACN/12.5 mM NH_4HCO_3. 2-Sulfobenzoic acid cyclic anhydride was prepared at a concentration of 2 mg/ml in dry tetrahydrofuran prior to use. A volume of 2 μl of the sulfonation reagent was mixed with 2 μl of the peptide extract, briefly vortexed, and reacted for 5 minutes at room temperature.

Matrix-Assisted Laser Desorption/ionization TOF/TOF Mass Spectrometry

The Applied Biosystems 4700 Proteomics Analyzer with TOF/TOF optics (Medzihradszky *et al*, 2000) was used in this study for reflectron analysis and MALDI MS/MS applications (Applied Biosystems, Framingham, MA). The mass spectrometer uses a 200-Hz frequency tripled Nd:YAG laser operating at a wavelength of 355 nm. For MS/MS, ions generated by the MALDI process were accelerated at 8 kV through a grid at 6.7 kV into a short, linear, field-free drift region. In this region, the ions pass though a timed-ion-selector device that is able to select one peptide from a mixture of peptides at different m/z values for subsequent fragmentation in the collision cell. After a peptide at a given m/z was selected by the timed-ion-selector, it passes through a retarding lens where the ions are decelerated and

then passes into the collision cell, which was operated at 7 kV. The collision energy, defined by the potential difference between the source and the collision cell, was 1 kV. Inside the collision cell, no collision gas was provided. After the collision, the ions are accelerated in the second source region at 15 kV, passed through a second, field-free, linear drift region, into the reflector, and finally to the detector. The detector amplifies and converts the signal to an electric current, which is observed and manipulated by a PC-based operating system. For the reflector mode, the operation of the instrument is far simpler. After the MALDI process generates the peptide ions, the latter are accelerated at 20 kV through a grid at 14 kV into the first, short, linear, field-free drift region. After this point, the rest of the instrument can be treated as a continuation of this region until the ions enter the reflector and then reach the detector, where as before, the signal at the detector is amplified and converted to electrical current.

The matrix solution was prepared as a 7 mg/ml α-cyano-4-hydroxycinnamic acid solution in 50 % ACN containing 0.1 % TFA. A volume of 1 µl of the sulfonated peptide was mixed with 1.5 µl matrix solution, vortexed, and 0.5 µl of the mixture was spotted on a 192-well stainless steel target plate (500 fmol peptide). The samples were allowed to air-dry at room temperature and were then inserted into the mass spectrometer and subjected to MALDI MS analysis. Prior to analysis, the mass spectrometer was externally calibrated with a mixture of Angiotensin I, Glu-fibrino-peptide B, ACTH (1-17), ACTH (18-39). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrino-peptide. MS and MS/MS data were further processed using DataExplorer 4.0 (Applied Biosystems, Framingham, MA) or by manual interpretation.

Results and Discussion

TOF/TOF analysis of native and derivatized peptides

Table 3.3. Overview of the peptide sequences and their masses (monoisotopic values) used in this case study

Peptide fragment	Sequence	Theoretical masses (Da) Experimentally observed masses (Da) Unmod.[b]	Guanid.[c]	Modif.[d]
α-S2 cas (A96-K106)	ALNEINQFYQK	1366.695 1367.643	1408.695 1409.676	1592.695 1591.270[a]
α-S2 cas (N174-K180)	NRLNFLK	903.536 904.512	945.536 946.518	1129.536 1130.490
α-S1 cas (Y106-R115)	YLGYLEQLLR	1266.705 1267.699	1266.705 1267.639	1450.705 1449.274[a]
α-S1 cas (E140-K147)	EGIHAQQK	909.474 910.530	951.474 952.538	1135.474 1136.516
α-S2 cas (T197-K203)	TVYQHQK	902.468 903.521	944.468 945.531	1128.48 1129.508
α-S2 cas (L168-R175)	LTEEEKNR	1017.516 1018.579	1059.516 1060.582	1243.516 1244.571
α-S1 cas (H23-R37)	HQGLPQEVLNENLLR	1758.945 1760.010	1758.945 1759.979	1942.945 1943.955

α-S2 cas (F189-K196)	FALPQYLK	978.561 979.640	1020.561 1021.519	1204.561 1203.248[a]
HHC (T40-K53)	TGQAPGFTYTDANK	1469.686 1470.723	1511.686 1512.692	1695.686 1694.529[a]
Synthetic	ELAQYNVEVHPYTVRK	1945.013 1945.971	1987.013 1987.906	2171.013 2171.638
Synthetic	APWFHHQNGK	1220.591 1221.361	1262.591 1263.379	1446.591 1447.345
Synthetic	QAQVYPNRFPLWK	1645.880 1646.790	1687.880 1688.574	1871.880 1872.549
α-S1 cas (H95-R105)	HIQKEDVPSER	1336.681 1337.745	1378.681 1379.783	1562.681 1563.781
HHC (T28-K39)	TGPNLHGLFGRK	1295.717 1296.773	1337.717 1338.715	1521.717 1522.806

The numbering in parentheses is according to the numbering of the α-casein (α-cas) and horse heart cytochrome c (HCC) sequences in the latest Swiss-Prot release. [a] ions observed in negative mode reflectron analysis, [b] mass of the unmodified peptides, [c] mass of the guanidinated peptides, [d] mass of the guanidinated and sulfonated peptides (modified). Underscored peptide sequences could be deduced from the MS/MS spectra of the modified peptides.

Peptides resulting from proteolytic digests with endoprotease LysC or trypsin were separated by HPLC. Individual fractions were analyzed by MALDI MS (reflectron mode) and judged to be pure if they contained only one or two major fragments. A summary of the peptide fragments used in these experiments is given in Table 3.3. The numbering of the amino acids of the peptide fragments is according the latest Swiss-Prot release (version 42.10). Derivatization of the individual peptide fragments was performed as described before. After isolation, 50 pmol of RPLC-purified peptide was guanidinated and desalted on a Prosorb-device. After extraction and assuming 100% recovery, 10 pmol of the peptide was sulfonated using 2-sulfobenzoic acid cyclic anhydride as the derivatization reagent rather than chlorosulfonylacetyl chloride, which is known to hydrolyze quickly with water. In all experiments, 500 fmol of the derivatized peptide was applied on the MALDI stainless steel probe. The minimum amount of protein required to obtain good quality MS/MS spectra varies from 10 pmol to 200 fmol, as recently reported (Lin *et al*, 2003). Fragmentation spectra were obtained from approximately 5000 laser shots of 500 fmol of the peptide in the metastable decomposition mode (gas off) and with the collision energy set at 1 keV. Under these experimental conditions, the major peaks in MS/MS spectra typically correspond to y- and b-series ions.

MALDI reflectron analysis of a fraction from the endoprotease LysC digest of α-casein indicated the presence of two peptides: ALNEINQFYQK (Ala96-Lys106, α-S2 casein) and NRLNFLK (Asn174-Lys180) at respectively 1367.643 and 904.512 Da. Upon guanidination (+ 42 Da) and sulfonation (+ 184 Da) the smallest fragment was clearly observed at 1130.490 Da but the modified, larger, fragment was not detected in positive ion mode. Analysis in negative mode showed the two fragments at 1128.228 and 1591.270 Da (Figure 3.8a). MS/MS analysis, in the positive mode, on the smaller precursor with m/z 1130.49, yielded a complete y-ion series (Figure 3.8c). The precursor ion first loses the sulfonation label (-184 Da) followed by formation of a complete series of y-ions. By simple manual calculation of the differences between the adjacent y-ion fragments the complete

sequence of the peptide could easily be determined. Fragment ions y_6 and y_7 are accompanied by their (y_i –17) satellite ions, most likely formed by neutral loss of ammonia from the internal arginine as discussed below. MS/MS analysis of the unmodified precursor from the larger fragment (m/z 1367.643) yielded a complex fragmentation pattern including incomplete series of b-, a-, and y-ions several internal fragment and immonium ions (Figure 3.8b). Even though it is possible to derive sequence information from such MS/MS spectra in some cases, the complexity of the fragment spectra often prevents an unambiguous sequence determination. After derivatization, the largest peptide was not observed in positive mode analysis. However, setting the first TOF analyzer to select the theoretical 1593.27 precursor (1591.27 + 2 H^+), followed by fragmentation, also yielded the complete y-ion series (Figure 3.8d). The same phenomenon was observed during analysis of the tryptic peptide fragment YLGYLEQLLR (Tyr1065-Arg115) from α-S1 casein (P02662) (theoretical mass 1266.705 Da). As this peptide contains no Lys its mass remained unchanged after guanidination. Upon sulfonation, the expected mass (1451.70 Da) was not observed in positive mode MS analysis but was detected as its deprotonated form in negative mode (Table 3.3). However, selection of the protonated ion (1451.27 Da) in positive mode fragmentation analysis yielded a spectrum containing exclusively y-ions (results not shown).

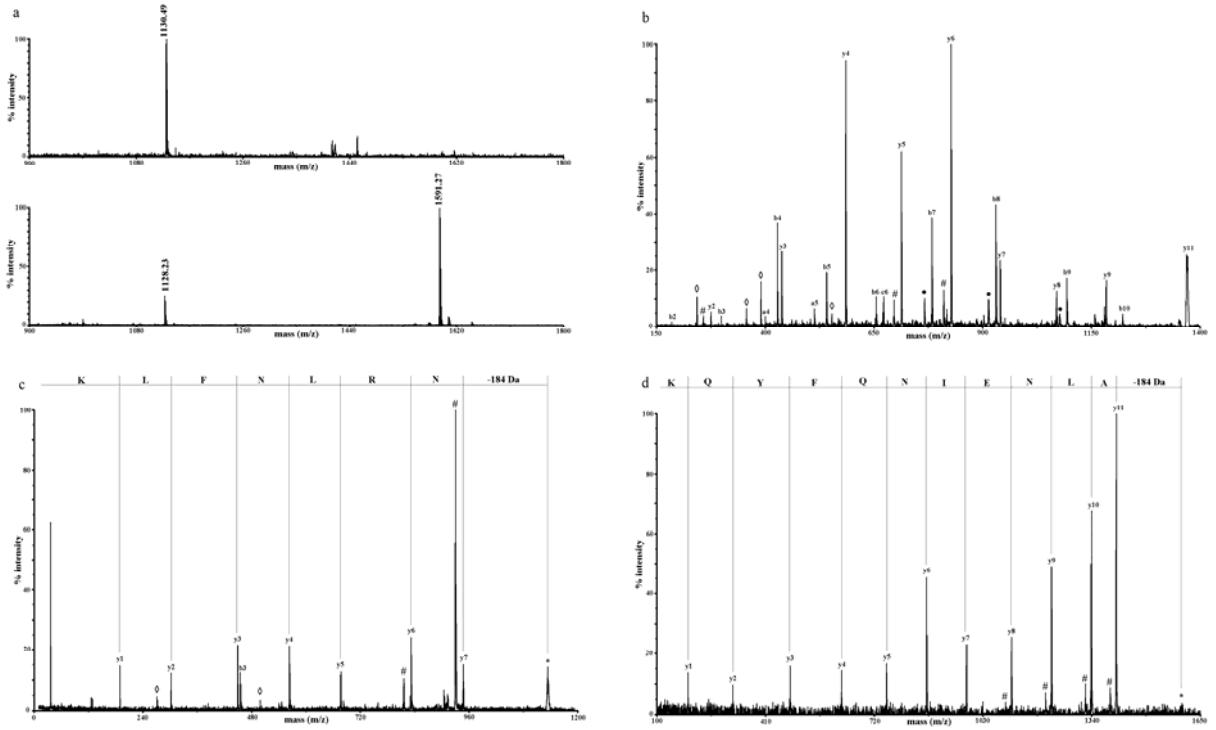


Figure 3.8. a) The positive (upper) and negative (lower) MALDI reflectron MS spectrum of a RPLC fraction from α-casein cleaved with endoprotease LysC. b) MALDI MS/MS spectrum of the m/z 1367.64 ion of the unmodified peptide. c) MALDI MS/MS spectrum of the m/z 1130.49 ion of the derivatized peptide. d) MALDI MS/MS spectrum of the m/z 1593.27 ion of the derivatized peptide (positive mode). In each spectrum only a-, b- and y-ions are labeled. The derivatized precursors are labeled with *. Where appropriate, (y-17)-ions (#), (b-17)-ions (•) and internal fragment ions (◊) are shown as indicated.

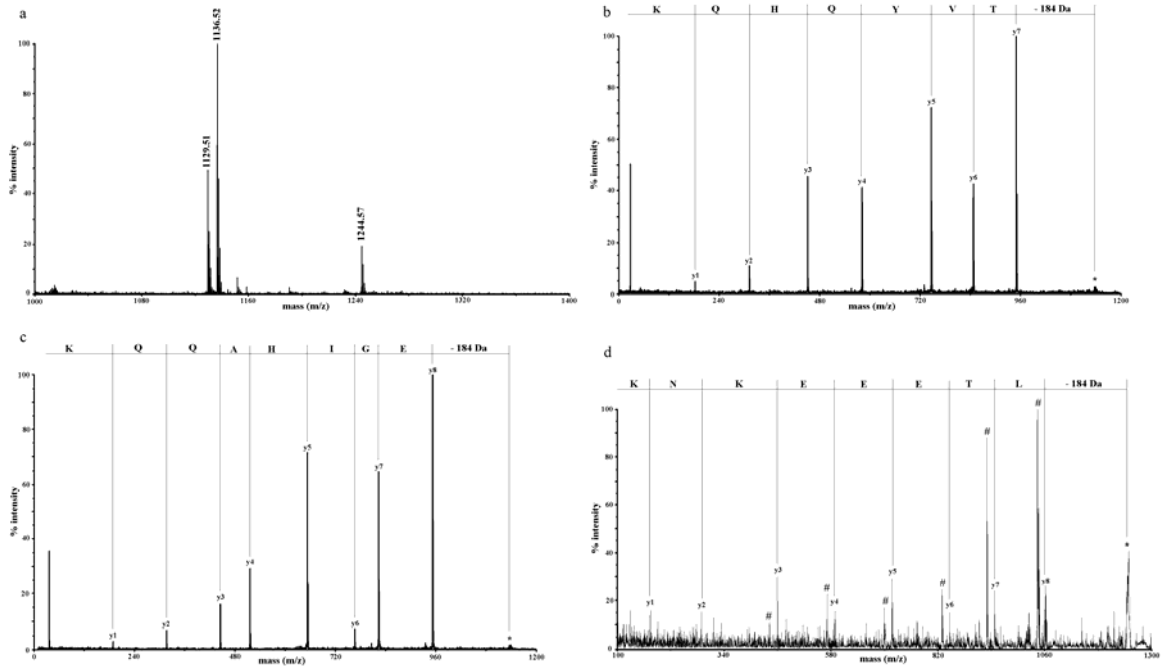


Figure 3.9. a) MALDI reflectron MS spectrum (positive mode) of a tryptic α-casein fraction after derivatization. Panels b), c) and d) show respectively the MALDI MS/MS spectra of the modified peptides at m/z 1129.51, m/z 1136.52 and m/z 1244.57. The loss of the sulfonation label (-184 Da) is indicated and (y-17)-ions are labeled with #.

Figure 3.9a shows the analysis of a second tryptic fraction wherein three components were observed: two major ones, EGIHAQQK (Glu140-Lys147, α-S1 casein) and TVYQHQK (Thr197-Lys203, α-S2 casein), and a minor one, LTEEEKNR (Leu168-Arg175, α-S2 casein) at respectively 910.530, 903.521 and 1018.579 Da (Table 3.3). After guanidination and N-terminal sulfonation the masses of all fragments increased respectively 42 and 184 Da (Figure 3.9a). Fragmentation of the underivatized peptides yielded complex fragmentation patterns consisting of N-terminal a- and b-ions, C-terminal y-ions and internal fragment ions. Again, these spectra would be difficult to interpret *de novo*. Fragmentation of the completely derivatized major components resulted in two complete y-ions series (Figure 3.9b & c) whereas fragmentation of the minor fragment precursor (m/z 1244.57), being approximately 5-fold lower in intensity (Figure 3.9a & d), resulted in a more complicated fragment spectrum. The y-ion series was evident in the spectrum, but all y-ions, except for y_1 and y_2, were accompanied with (y-17)-ions approximately twice as large as their corresponding y_i ions. The formation of this second series is due to the presence of an internal homoarginine which is known to lose ammonia (-17 Da). The loss of neutral molecules such as ammonia has also been observed during MALDI-analysis of peptides containing internal Arg residues.

Peptides having internal amino acids known to induce specific fragmentation

Two tryptic peptide fragments, HQGLPQEVLNENLLR (His23-Arg37, α-S1 casein) and FALPQYLK (Phe189-Lys196, α-S2 casein), were subjected to modification and subsequent MALDI TOF/TOF analysis. Upon guanidination and sulfonation of the second peptide fragment, FALPQYLK, again the peptide was not observed during MALDI reflectron analysis in the positive mode but only as its deprotonated ion in the negative mode (1203.248 Da, Table 3.3). Both the unmodified (m/z 979.640) and the completely modified peptide (m/z 1205.25, not observed in positive mode analysis) were fragmented. As described above, fragmentation of the native peptide yielded a mixture of incomplete b- and y-ion series

(wherein the y_5-ion is the most abundant) and a number of internal fragment ions, whereas fragmentation of the modified peptide (m/z 1205.25) yielded the complete y-ion series (Figure 3.10a). As for the underivatized peptide, a preferential cleavage at the N-terminal side of Pro was observed, resulting in the presence of a dominant y_5-ion and a very weak, although detectable, y_4-ion. Since fragment His23-Arg37 contains no lysine its mass remained unchanged upon guanidination (1759.979 Da). In the MS/MS fragment spectrum of the underivatized peptide, three major y-ions were observed, y_{11}, y_8 and y_4, indicating preferential fragmentation, respectively N-terminal of Pro5 and C-terminal of Glu7 and Glu11. Fragmentation of the sulfonated precursor ion (1943.955 Da) yielded the complete y-ion series (y_1-y_{15}) wherein the y_{11}-ion was the most dominant fragment ion (Figure 3.10b). Although the intensity of the y_{10}-ion was low the ion could be observed in the fragmentation spectrum. Preferential cleavage C-terminal of the Glu residues was not longer observed (Figure 3.10b). Upon complete derivatization, peptide TGQAPGFTYTDANK (Thr40-Lys53, horse heart cytochrome c) was only observed in negative mode analysis (1694.53 Da). Again, fragmentation of the hypothetical protonated precursor (1696.53 Da) yielded the complete y-ion series and indicated a preferential cleavage at the Ala-Pro bond. No preferential cleavage C-terminal of the Asp residue was observed, neither in the fragment spectra of the underivatized fragment nor in that of the sulfonated peptide.

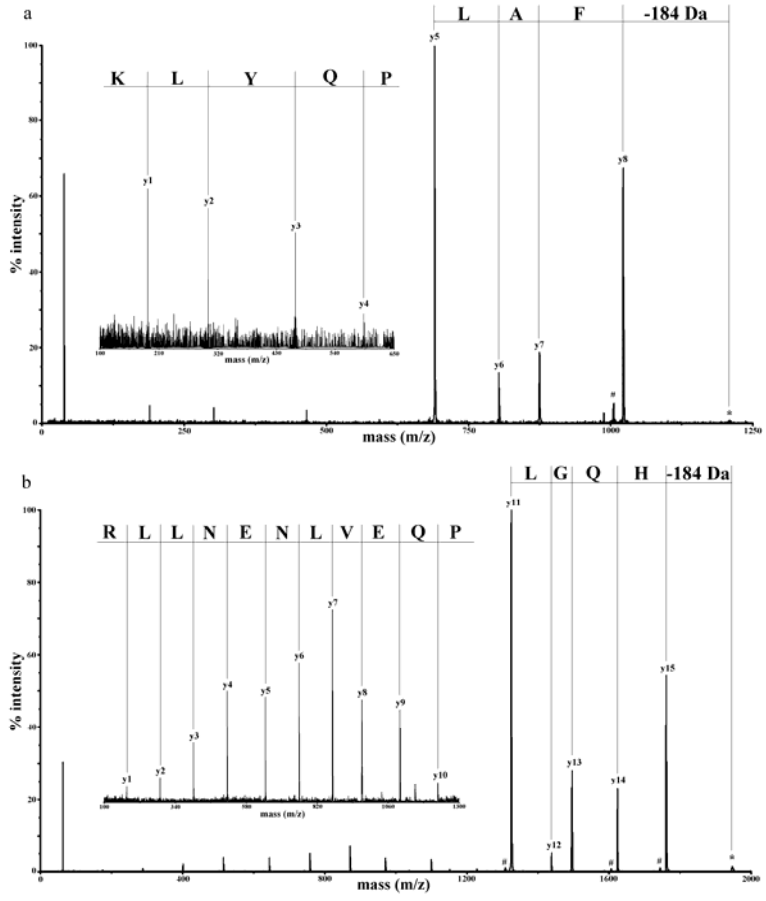


Figure 3.10. MALDI MS/MS spectrum of the modified tryptic peptides FALPQYLK (m/z 1205.25)(a) and HQGLPQEVLNENLLR (m/z1943.96) (b). The derivatized precursors are labeled with * and the loss of the sulfonation label (-184 Da) is indicated. The insets show an expanded view of the y-ion series obtained after preferential cleavage at the Xxx-Pro bond.

From recent studies it is known that peptides containing internal basic residues (Arg, homo-Arg or His) do not fragment as readily as typical tryptic fragments (Tabb *et al*, 2004). MALDI-TOF/TOF analysis of the underivatized synthetic peptide ELAQYNVEVHPYTVRK yielded an incomplete y-ion series, in which the y_6-, y_7-, y_8- and y_{15}-ion were the most abundant fragment ions (Figure 3.11b). The predominance of y_8 and y_{15} indicates a preferential cleavage of the Glu-Xxx bonds. The y_6-ion is due to preferential cleavage of the His-Pro bond whereas the y_7 results from a preferential cleavage of the Val-His bond. Fragmentation of the derivatized peptide (precursor m/z 2171.64) yielded the complete y-ion series accompanied with a more pronounced (y-17)-ion series due to the presence of Arg at the penultimate position. Some of these (y-17)-ions were also observed in the fragmentation spectrum of the underivatized peptide (Figure 3.11b). The y_7-, y_8-, and y_{15}-ions are no longer more pronounced in the spectrum, in contrast to the y_6-ion, indicating that the preferential His-Pro bond cleavage still occurs (Figure 3.11d). Although weaker, the y_1- to y_5-ion series could still be observed. The influence of the presence of an internal His on the fragmentation was also observed during analysis of the synthetic peptide APWFHHQNGK. MALDI-TOF/TOF analysis of the underivatized fragment (m/z 1221.36) showed a mixture of b-, a- and y-ion series. The most dominant ions are the y_5- and y_6-ions, indicating preferential cleavage at the Xxx-His bonds (Figure 3.11a). In contrast to previous observations, the y_9-ion, indicating a preferential Ala-Pro bond cleavage, was not dominantly present in this fragmentation spectrum. Fragmentation of the derivatized peptide (precursor 1447.34 Da) yielded a completely different fragmentation spectrum with the y_9-ion being the most dominant fragment ion in the y-ion series (Figure 3.11c). The y_5- and y_6-ions were only slightly more pronounced in this y-ion series.

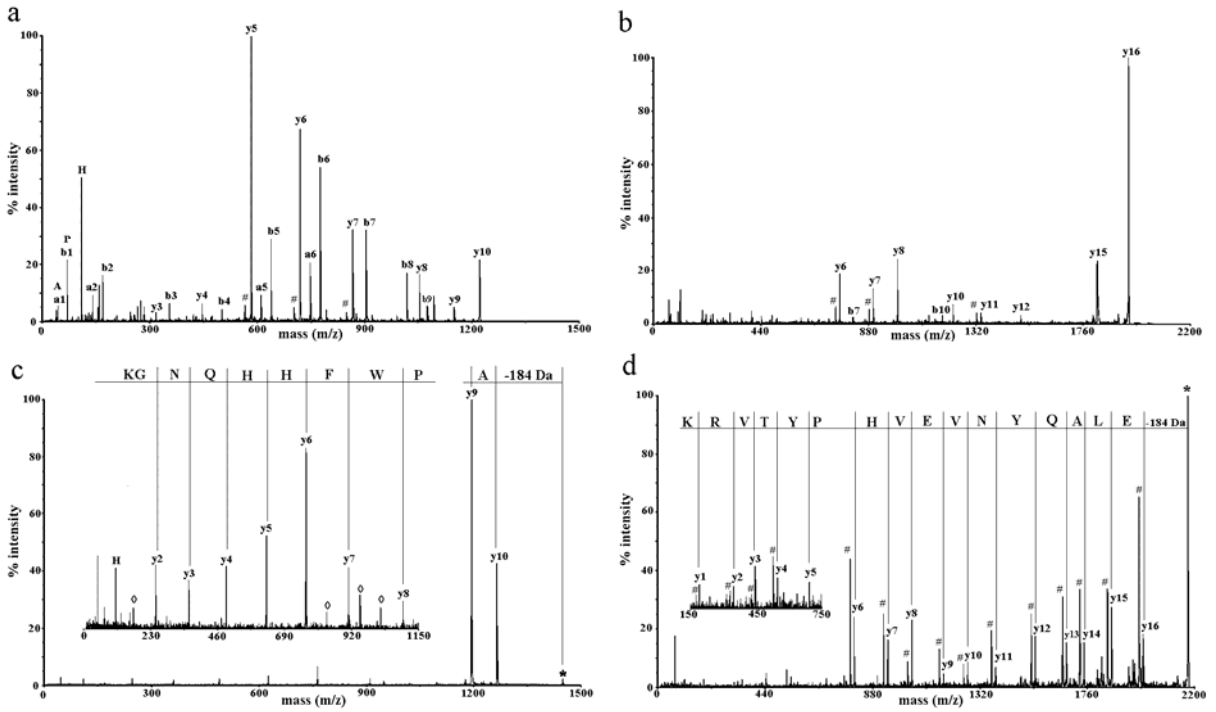


Figure 3.11. MALDI MS/MS spectra of the underivatized (a) and modified (b) peptide APWFHHQNGK. MALDI MS/MS spectra of the underivatized (b) and modified (d) peptide ELAQYNVEVHPYTVRK. The insets show an expanded view of the y-ion series obtained after preferential cleavage at the Xxx-Pro bond. In each spectrum only a-, b-and y-ions are labeled. The derivatized precursors are labeled with *. The loss of the sulfonation label (-184 Da) is indicated and, where appropriate, (y-17)-ions (#) and internal fragment ions (◊) are shown as indicated.

TOF/TOF analysis of the underivatized synthetic peptide QAQVYPNRFPLWK yielded a rather complex fragmentation spectrum wherein no dominant fragment ions were observed. The TOF/TOF fragment spectra obtained from the sulfonated homoarginine-terminated peptide is shown in Figure 3.12a. The y-ion fragment series y_6-y_{13} is accompanied with the more intense (y_i -17)-ion series due to the loss of ammonia from the internal Arg. Due to preferential fragmentation at the Xxx-Pro amide bonds the spectrum contains predominantly the y_4- and the (y_8 –17)-ions, whereas the y_3- and the y_7-ions are not or hardly observed. Due to the presence of a consecutive Pro-Asn sequence in the peptide care must be taken to interpretate this spectrum for assigning the correct sequence (Asn, 114 Da = [Pro, 97 Da + ammonia, 17 Da]).

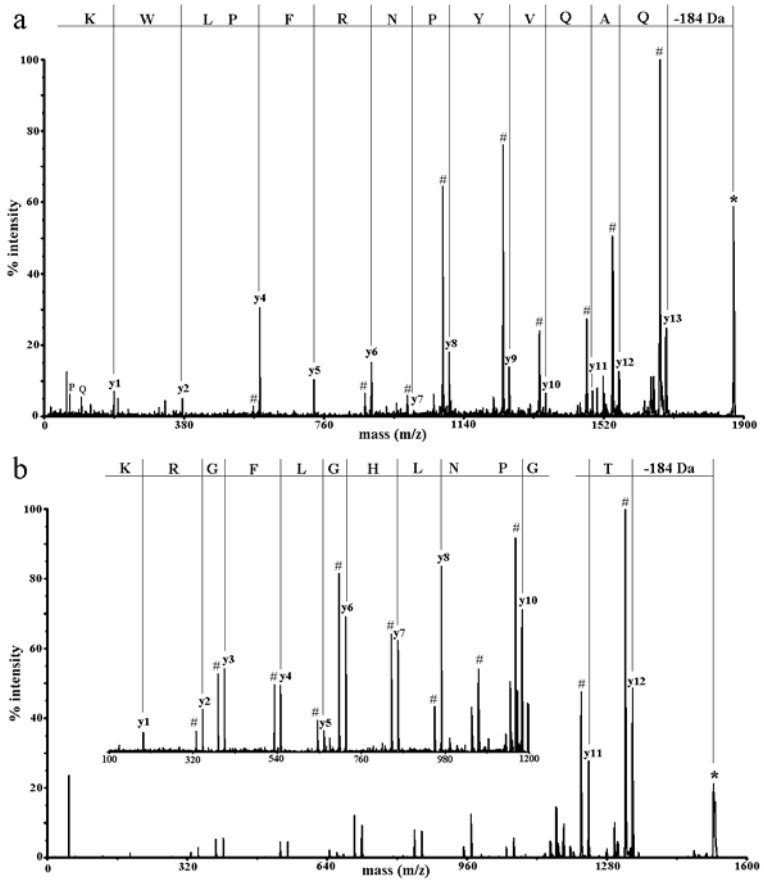


Figure 3.12. MALDI MS/MS spectrum of the modified peptides QAQVYPNRFPLWK (m/z 1872.55) (a) and TGPNLHGLFGRK (m/z 1522.81) (b). The derivatized precursors are labeled with * and the loss of the sulfonation label (-184 Da) is indicated. The inset shows an expanded view of the y-ion series obtained after preferential cleavage at the Xxx-Pro bond. (y-17)-ions (#) are shown as indicated.

The same problem was encountered during TOF/TOF analysis of peptide TGPNLHGLFGRK (Thr28-Lys39, HHC). Due to the presence of Arg as the penultimate amino acid, the fragmentation spectrum of the derivatized peptide (m/z 1522.81) was dominated by the (y-17)-ion series (y_2-y_{12}). Together with the presence of a consecutive Pro-Asn (Δm=17 Da) in the sequence this might lead to an incorrect sequence assignment (Figure 3.12b). Fragmentation of the underivatized tryptic fragment HIQKEDVPSER (His95-Arg105, α-S1 casein) yielded a fragmentation spectrum with a b- and y-ion series in which the y_4 and y_5 were the most intense fragment ions (Figure 3.13a), indicating a preferential cleavage of the Val-Pro and the Asp-Val peptide bond respectively. In the fragmentation spectrum of the derivatized peptide (m/z 1563.78) the same ions were observed as the most intense fragment-

ions (Figure 3.13b). Apart from the complete y-ion series, a more pronounced (y_i -17)-ion series was observed due to the presence of an internal homo-Arg in the peptide.

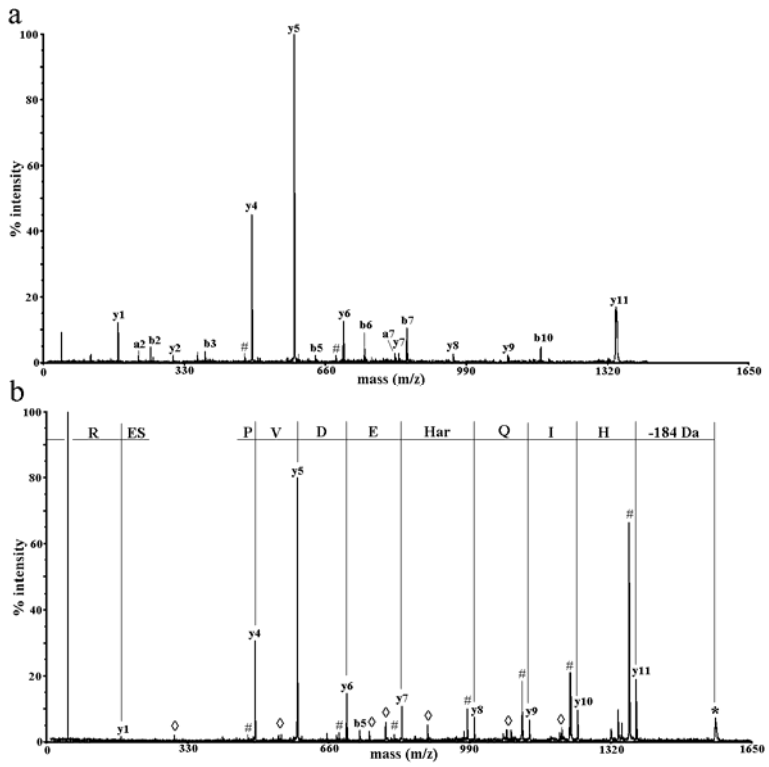


Figure 3.13. MALDI MS/MS spectra of the underivatized (a) and modified (b) peptide HIQKEDVPSER. In each spectrum only a-, b- and y-ions are labeled. The derivatized precursor is labeled with *. Where appropriate, (y-17)-ions (#) and internal fragment ions (◊) are shown as indicated.

De novo sequence analysis

It is known that the introduction of a strong acidic group at the N-terminus of tryptic peptides facilitates protonation of backbone amide bonds, leading to extensive structure-specific fragmentation under PSD MALDI and electrospray MS/MS conditions (Keough *et al*, 1999; Bauer *et al*, 2000; Keough *et al*, 2000b). In the present study, the effect of the N-terminal sulfonation on the CID TOF/TOF fragmentation pattern was investigated. The sulfonation reaction results in the modification of both the N-termini and the ε-amino groups of lysine-containing peptides. For unprotected Lys-terminated peptides this will result in the formation of disulfonate derivatives. The formation of such derivatives was previously shown to be undesirable because they exhibit poor sensitivity in the positive-ion mode and relatively poor fragmentation under negative-ion analysis conditions. Negative-ion PSD spectra of several Lys-containing disulfonate derivatives showed both low-product ion yields and complex fragmentation patterns containing b- and y-type ions linked to the sulfonate group (Keough *et al*, 2000b). Therefore, this approach requires a preliminary modification of the ε-amino group of lysine residues. It was demonstrated that, following guanidination of lysine ε-amines, introduction of sulfonic acid groups to tryptic peptides is possible solely at the N-terminus. The guanidination, leading to C-terminal lysines being converted into homoarginines (+ 42 Da), can selectively and quantitatively be performed with *O*-methylisourea at high pH and does not affect the peptide amino terminus or other side groups. The resulting homoarginine has an even higher pKa than arginine, and it has previously been shown that a rise of the basicity of lysines increases their relative abundances in MALDI mass spectra by enabling enhanced charge retention (Beardsley *et al*, 2000; Hale *et al*, 2000).

Furthermore, it allows unambiguous differentiation of Lys (128.18 Da) and Gln (128.13 Da) by mass spectrometry.

CID experiments on several underivatized peptides with a prototype MALDI-TOF/TOF mass spectrometer indicated that the nature of the MALDI-CID spectra is quite dependent upon the amino acid composition of the peptide, the peptide size, the matrix, and the collision gas used. Different combinations of MALDI matrix and collision gas determine the amount of internal energy deposited by the MALDI process and the CID process. In our study, all experiments were performed using α-cyano-4-hydroxycinnamic acid (α-CHCA), the most commonly used matrix for peptide analysis. α-CHCA is a relative hot matrix and is known to produce unimolecular decomposition (also called PSD) which is known to vary with the amino acid sequence. In a preliminary study, fragmentation spectra resulting from high- and low-energy CID were compared (Sumpton *et al*, 2003). The authors concluded that the difference in fragmentation and the effect on database search results was surprisingly small. The major difference observed is the presence of high-energy fragment ions (w-ions) in the high-energy CID spectra of some peptides. When the collision induced dissociation mode (gas on, collision energy 0.5 to >3.5 keV) is utilized, a larger number of low molecular weight fragments (immonium ions, internal fragments) have been observed (Walker *et al*, 2003). However, it has also been shown that the use of high-energy CID results in a loss of sequence information as the y-ion abundance decrease at both higher gas pressure and higher collision energy (Campbell, 2003). Therefore, we performed all fragmentation experiments with the collision energy set at 1 keV and no gas in the collision chamber (low-energy CID).

In the present case study, we used a set of peptides purified from proteolytically cleaved proteins with known sequences, rather than using test peptides. We found this set to be more representative for 'real life' samples as some of the peptides had internal basic residues resulting from missed cleavages, some RPLC fractions contained mixtures (non-stoichiometric amounts), and most peptides had one or more amino acids in their sequence known to induce specific fragmentation. TOF/TOF analysis of underivatized peptides typically results in complex fragment spectra, containing incomplete a-, b- and y-ion series (Figure 3.11a & b, 3.13a). The complexity of the spectra is often increased by the presence of internal fragment ions (Figure 3.8b). Similar complex fragmentation spectra have also been observed in fragmentation studies of singly charged precursor ions using a MALDI quadrupole TOF instrument (Wattenberg *et al*, 2002). The guanidinated lysine-terminated peptides showed better positive ion MALDI response than native lysine-terminated peptides. From MS/MS experiments of homoarginine-terminated peptides it was apparent that guanidination does not enhance peptide fragmentation efficiency (results not shown). This is consistent with earlier findings that increasing the basicity of the peptide increases the internal energy required for fragmentation (Dongré *et al*, 1996).

Cleavage of amide bonds under low-energy collision activation conditions is generally thought to be initiated by migration of the charge from the initial site of protonation (e.g. the N-terminal amino group or the side chains of basic amino acids such as arginine, lysine, and histidine) to amide carbonyl oxygen along the peptide backbone. This 'mobile proton' termed hypothesis is one of the central tenets in peptide fragmentation mechanisms (Wysocki *et al*, 2000). Fragmentation of the peptide amide bond then occurs by neighboring group attack from an adjacent nucleophilic amide carbonyl moiety to yield complementary b- and/or y-ions (a charge-directed process). Most ESI MS/MS sequencing experiments use doubly charged protonated peptides because they fragment readily. Unfortunately, doubly protonated molecules are not formed in high yield under MALDI conditions and, therefore, these labile

ions are not available for sequencing studies using this technique. The addition of a sulfonic group to the N-terminus allows to create a peptide with an extra proton, but not a double charge. This strategy concedes that the basic C-terminal residue will be protonated under MALDI conditions. The strong acid is chosen such that it would remain deprotonated under MALDI conditions, counter-balancing the C-terminal positive charge. The additional proton, required to ionize the peptide for MS analysis, is more or less free to randomly protonate amide bonds, as the most basic site in the molecule is already occupied (Keough *et al*, 2003).

After guanidination and sulfonation a contiguous series of y-ions in all of the fragmentation spectra was observed. The y-ion series could easily be interpreted (manually or by using an algorithm) facilitating *de novo* sequencing. Although there is a dramatic difference in speed, cost and sensitivity, the ease by which *de novo* sequencing can be performed by MALDI TOF/TOF analysis of derivatized peptides can be compared with classical Edman degradation analysis. In all fragment spectra an initial loss of the sulfonic acid derivative is observed (Δm = 184 Da). A significant loss of the derivative was also observed by others using vacuum MALDI MS when the analyses were conducted using α-CHCA as matrix (Keough *et al*, 1999). The results indicated that some of the sulfonic acid-derivatized peptides had poorer positive-ion sensitivity than the corresponding native peptides and, after derivatization, four of the peptides were no longer observed in positive mode reflectron analysis. These fragments could be detected as their deprotonated ions when the analysis was performed in the negative mode (Table 3.3). However, selection of the corresponding protonated precursor ion for TOF/TOF analysis (positive mode) results in the formation of complete series of y fragment ions (Figure 3.8d, 3.10a). Most likely the protonated precursor is not detected due to metastable decomposition. Keough *et al* already noticed a decreased intensity of sulfonated peptides in the positive ion mode, compared to the negative ion mode and, apparently, some peptides show no signal above the noise level. However, the metastable decomposition yields excellent MS/MS spectra in the positive ion mode, as illustrated in this work.

Amino acid dependent specific fragmentation

For many years, low-energy collision induced dissociation (CID) has been the activation method of choice in the attempt to identify proteins by means of gas phase fragmentation of one or more of their peptides (Mann *et al*, 2001). However, many factors can influence peptide dissociation in the gas phase including the nature of the residues present, the charge state of the precursor ion, and the size and conformation of the peptide. In some instances, some of these factors can lead to enhanced or specific cleavage at certain peptide bonds, limiting the information obtained from the MS/MS spectrum. It has long been recognized that mass spectrometric peptide sequencing techniques would be improved if the relative fragmentation efficiencies of the cleavage sites in a peptide could be predicted. Such studies included the identification of residues that enhance specific ion fragmentation pathways. Most of these studies have focused on fragment spectra resulting from MS/MS of doubly protonated tryptic peptides. In this context, a highly efficient fragmentation of backbone amide bonds on the N-terminal side of proline residues has been described by several authors (Loo *et al*, 1993; Breci *et al*, 2003). Other investigators have elucidated dissociation pathways that are promoted by acidic residues (Yu *et al*, 1993). Preferential cleavages at Pro and His (Tabb *et al*, 2003) and at the acidic residues Asp and Glu (Bailey *et al*, 2003) have been indicated. The selective cleavage of protonated peptides, derivatized with a fixed-charge, containing internal Asp or His residues has been reported (Gu *et al*, 2000; Tsaprailis *et al*, 2004). The influence of the basic residue content on the fragmentation

behavior has also been studied (Gu *et al*, 1999; Tsaprailis *et al*, 2000; Huang *et al*, 2002; Tabb *et al*, 2004).

Sophisticated algorithms (e.g. SEQUEST, Mascot) have been developed for identifying proteins from peptide MS/MS data. Peptides are hereby identified by correlating the uninterpreted MS/MS spectra with simulated product ion spectra using a relative simple set of previously defined parameters regarding the expected fragmentation behavior of protonated peptide ions. However, the current understanding of the fragmentation mechanisms is still insufficient to ensure a high correlation between theoretically predicted MS/MS spectra and experimental results. Furthermore, the fact that selective/enhanced cleavage does not always occur might have discouraged investigators from including them in the existing computer-based interpretation of peptide MS/MS spectra. Therefore, any improved understanding of the enhancement or absence of certain fragment ions in MS/MS experiments provides additional and better predictive rules for the interpretation of peptide MS/MS spectra. Recently, several groups have systematically examined databases of tryptic MS/MS spectra to evaluate the effects of specific residues or charge states on the peptide fragmentation and fragment ion intensity (Yu *et al*, 1993; Kapp *et al*, 2003; Nesvizhskii *et al*, 2003; Shütz *et al*, 2003; Tabb *et al*, 2003)

In a preliminary study, in which MALDI-TOF/TOF was used to produce MS/MS spectra of singly charged underivatized peptides, dominant product ions resulting from Xxx-Pro fragmentation were observed (Walker *et al*, 2003). These authors studied the effects of the MALDI matrix and the CID gas on the fragmentation efficiency of larger peptides and small proteins. They observed that fragmentation becomes more sequence specific, with a cleavage selectively at the N-terminus of Pro as well as at the C-terminus of the acidic residues, as the molecular weight of the peptide is increased (> 4000 Da). MS/MS spectra generated by MALDI quadrupole TOF mass spectrometry of singly charged peptides also showed preferential fragmentation at the N-terminal bond of Pro and at the C-terminal bond of the acidic residues Asp and Glu (Wattenberg *et al*, 2002). The same observation was made during CID analysis of singly charged peptide ions using an in-house assembled MALDI-quadrupole ion trap (Zhang *et al*, 2003). Several of the peptides used in our study contained Pro and/or acidic residues and some had, due to missed cleavages, internal Lys or Arg (Table 3.3). TOF/TOF analysis of the underivatized peptides was performed to explore any known or novel preferential fragmentation patterns. The influence of the sulfonation reaction on the observed preferential fragmentation, and hence the possibility to obtain complete y-ion series, was determined by comparing the TOF/TOF spectra before and after derivatization.

Fragmentation of peptides containing proline residues

When the ionizing proton is freely ‘mobile’, enhanced cleavage is commonly observed at Xxx-Pro bonds, presumably due to higher local proton affinity of the proline imide bond compared to a conventional amide bond. During PSD experiments with derivatized peptides Keough *et al* observed enhanced fragmentation at Pro residues and a reduced abundance in fragmentation at the C-terminal side of Pro (Keough *et al*, 2000b). Therefore, a Pro residue near the N-terminus of a peptide can limit the amount of sequence information that can be derived from the peptide. In our study, TOF/TOF analysis of peptides containing an internal Pro indicated that y-ions resulting from cleavage on the N-terminal side of Pro are enhanced while y-ions resulting from cleavage on the C-terminal side of Pro are less abundant or almost completely depleted (Table 3.3). MS/MS analysis of Pro-containing underivatized peptides indicated a preferential cleavage at the Xxx-Pro bond except for peptides APWFHHQNGK

and TGPNLHGLFGRK. MS/MS analysis of the underivatized fragment QAQVYPNRFPLWK, having two internal Pro residues, indicated a cleavage only at the second Pro (y_4) (Figure 3.13a). In contrast, upon complete derivatization, all peptides showed preferential cleavage at all Xxx-Pro bonds, with the exception of the peptide TGPNLHGLFGRK from HHC (Figure 3.12b). In a recent study, the unusual fragmentation effect in tandem mass analysis of doubly charged Pro-containing peptides was systematically analyzed (Breci *et al*, 2003). From this study it became apparent that ions formed at positions C-terminal to Pro, where Xxx is any amino acid, produce low Pro-Xxx cleavage. In contrast, preferential cleavage N-terminal to Pro occurred with strong ions formed at Xxx-Pro when Xxx was His, Asp, Ile, Leu, and weak or no ions formed when Xxx was Gly or Pro as is the case for the peptide TGPNLHGLFGRK.

Fragmentation of peptides with internal basic residues (Arg/homo-Arg/His)

Peptide resulting from incomplete protein digests show differences in their fragmentation behaviour due to the presence of internal basic residues. This type of residues can affect fragmentation in a variety of ways. Their presence in a peptide raises the minimum energy requirement for fragmentation (Dongré *et al*, 1996). Their side chains may be the sites in a peptide for which protons show the greatest affinity, in some cases resulting in proton sequestration and enabling charge-remote fragmentation mechanisms (Gu *et al*, 2000; Huang *et al*, 2002). The protonated side chains of His and Arg may also cleave adjacent amide bonds to form b-ions of non-oxazolone structure (Wysocki *et al*, 2000). Tabb *et al* recently described a statistical analysis of the influence of basic residue position and identity on fragment ion intensities in tandem mass spectra of non-tryptic peptides (Tabb *et al*, 2004).

After derivatization with the commercially available CAF-reagent incomplete y-ion series were observed during PSD experiments of peptides containing internal Arg (Hellman *et al*, 2002). In contrast to these results, we observed that TOF/TOF analysis of derivatized peptides with an internal Arg or homo-Arg results in the formation of a complete y-ion series. However, in all these spectra the y-ion series were accompanied with a more abundant (y_i – 17)-ion series due to the neutral loss of NH_3. Although this (y_i –17) series shows up in the fragment spectra until the basic residue is cleaved off (Figure 3.8c, 3.9d, 3.11d, 3.12a & b and 3.13b), its presence does not interfere with the sequence interpretation, so that the complete sequence of all the peptides could be determined (Table 3.3). However, as illustrated in Figure 5, the neutral loss (Δm = 17Da) might prevent a correct sequence interpretation for peptides having a Asn-Pro bond in their sequence (Δm Asn/Pro = 17 Da). The presence of a (y_i –17)-ion series, twice as large as the corresponding y_i-ions, was also observed by Keough *et al* during PSD analysis of derivatized peptides containing two arginine residues (Keough *et al*, 1999).

Recently, it has also been noted that product ion spectra may be influenced by the presence of histidine in a peptide (Huang *et al*, 2002; Tabb *et al*, 2004). Data mining of a set of peptide spectra showed that histidine is the amino acid residue most likely to give preferential cleavage at its C-terminal side in doubly protonated tryptic peptides (Tabb *et al*, 2003). In contrast, TOF/TOF analysis of several underivatized peptides containing internal His residues shows a preferential cleavage at Xxx-His (Figure 3.11a & b). More abundant y-ions were observed in the MS/MS spectra of EGIHAQQK (y_5), TVYQHQK (y_3), ELAQYNVEVHPYTVRK (y_7) and APWFHHQNGK (y_5 and y_6). This preferential cleavage was not, or less abundantly, observed in the fragment spectra of the derivatized peptides (Figure 3.9b & c, 3.11c & d).

Fragmentation of peptides containing acidic residues

Some MS/MS spectra of native peptides containing Asp or Glu residues are dominated by cleavage on the C-terminal side of the acidic residues. The fragment spectra of the underivatized peptides YLGYLEQLLR, ELAQYNVEVHPYTVRK, and HQGLPQEVLNENLLR were dominated respectively by the y_4-, y_8/y_{15}-ions (Figure 3.11b) and the y_4/y_8-ions, whereas the fragment spectrum of the native peptide HIQKEDVPSER showed a dominant y_5-ion (Figure 3.13a). In contrast, some of the peptides resulting from endoprotease Lys-C cleavage, or containing an internal Lys (ALNEINQFYQK, LTEEEKNR and TGQAPGFTYTDANK), did not show preferential cleavage at the acidic residues residues. Preferential fragmentation at Asp-Xxx bonds of underivatized peptides on a TOF/TOF instrument has also been observed (Lin *et al*, 2003). Tandem mass spectra of the derivatized peptides containing Asp or Glu show a more uniform fragmentation along the peptide backbone. Only one peptide, HIQKEDVPSER, showed a slightly more abundant cleavage C-terminal to Asp (y_5) after derivatization (Figure 3.13b). A more uniform y-ion series in the MS/MS of sulfonic acid derivatized tryptic peptides spectra using atmospheric pressure MALDI ion trap MS was also observed by others (Keough *et al*, 2001).

C-terminal cleavage at aspartic acid residues, a low-energy fragmentation pathway, was first observed in peptides containing Asp-Pro and Asp-Xxx bonds (Yu *et al*, 1993). Selective gas-phase cleavage at the peptide bond C-terminal of an Asp residue in protonated peptides has been investigated by several research groups. Recent investigations show that when the ionizing protons are not mobile, i.e. when they are sequestered by Arg, cleavage catalyzed by the acidic hydrogen of the Asp side chain becomes pronounced (Tsaprailis *et al*, 2000; Wysocki *et al*, 2000). This is supported by data for related peptides that have been derivatized to add a fixed charge to them instead of a proton; even those peptides that contain no added proton fragmented selectively at the aspartic acid (Gu *et al*, 2000). Inspection of the singly charged spectra of 10 tryptic peptides containing Arg plus Asp and/or Glu showed that all of them fragment selectively at the acidic residues (Wysocki *et al*, 2000). However, peptides containing lysine did not show the same trend, thus spectra of singly charged peptides that contain Lys and/or Asp/Glu showed no enhancement of cleavage at acidic residues (Gu *et al*, 1999). When protons in excess of the number of Arg are present, enhanced/selective cleavages of Asp-Xxx bonds are not expected (Gu *et al*, 2000). This is the case for the N-sulfonated peptides, as the introduction of a negative charge is counterbalanced by a second mobile proton. This proton is more or less free to randomly ionize the peptide backbone amide groups independently of the presence of acidic residues.

Conclusions

Although *de novo* sequencing of underivatized peptides using MALDI-TOF/TOF has recently been demonstrated (Yergey *et al*, 2002), the interpretation of fragment spectra from peptides originating from unknown proteins strongly depends on the use of automated search routines or on manual interpretation. TOF/TOF fragmentation analysis of underivatized peptides yields multiple, incomplete fragment ion series, which are often difficult to interpret. There is no guarantee that the information in such MS/MS spectra is sufficient for peptide identification. As proposed by Keough *et al*, the addition of a sulfonic acid group to the N-terminus of a peptide promotes efficient charge-site-initiated fragmentation of backbone amide bonds (Keough *et al*, 1999). Upon guanidination and sulfonation, all peptides resulting from a proteolytic cleavage with trypsin or endoprotease LysC digest, show complete y-ion series in their fragment spectra. By simple manual calculation of the differences between the

adjacent y-ion fragments, or by using suitable software, the amino acid sequence can readily be interpreted. In the preceding guanidination reaction, Lys is converted to home-Arg and therefore can easily be differentiated from Gln; unfortunately, Leu and Ile cannot be distinguished. As the mass of the parent peptide is known from the initial reflectron analysis, the final assignment can be verified. Furthermore, the use of a TOF/TOF instrument provides improved ion mass measurement accuracy (results not shown). After derivatization, some of the peptides were solely observed in negative ion mode analysis (Table 3.3). However, selection of the corresponding precursor in positive ion mode analysis allowed obtaining the full y-ion series. The difference in fragmentation behaviour between sulfonated and underivatized peptides has important consequences for protein identification via database searches. Search algorithms, such as FASTS and FASTF, which use multiple peptide sequences to identify homologous sequences in protein or DNA databases have recently been described (Mackey *et al*, 2002). Given the size and growth of the current databases, most proteins more than likely already have homologues in a database (Shevchenko *et al*, 2001). Database searching with MS-derived *de novo* peptide sequences allow proteomic identification of proteins from organisms whose genomes have not been sequenced.

MS/MS analysis of several underivatized peptides indicated a preferential cleavage at Xxx-Pro bonds and C-terminal of the acidic residues. These results are supported by previously described fragmentation mechanisms, supporting the mobile proton mode. In contrast to previously published results (Tsaprailis *et al*, 2004), a preferential fragmentation at several Xxx-His bonds was observed in the fragment spectra of the underivatized peptides. Upon guanidination and sulfonation, preferential fragmentation at His and the acidic residues was no longer observed in the MS/MS spectra. In contrast, preferential cleavage at Xxx-Pro was still observed and sometimes even more pronounced in the MS/MS spectra of the derivatized peptides. Although the y-ions resulting from the Pro-Xxx were sometimes less intense, the complete sequence could be determined *de novo* for most peptides. The 14 peptides constitute 150 amino acids, of which 142 could be determined (Table 3.3). The fragment spectra of derivatized peptides containing an internal Arg or homo-Arg all showed y-ion series accompanied with a more dominant (y_i –17)-ion series. As proteomic strategies are becoming increasingly reliant on the use of automated database search algorithms, incorporation of 'fragmentation rules', such as the observed preferential cleavage at Xxx-Pro peptide bonds, into the database search algorithms will aid in the development of more effective tools for high-throughput protein identification. Furthermore, the occurrence of 'non-sequence' specific ion fragments, such as the neutral loss of ammonia from peptides with internal Arg or homo-Arg, can be used to improve predictive models of peptide fragmentation for *de novo* sequence analysis.

3.1.3. Protein identification across species-boundaries

Different approaches have been proposed to correlate mass spectral data with the determination of the function of proteins. With the growing number of complete genomic sequences, powerful search algorithms such as MASCOT or SEQUEST allow the identification of proteins, in seconds. After the determination of the complete genome from *Haemophilus influenza* (Fleischmann *et al*, 1995) the pace of nucleotide sequence determination has dramatically increased. However, the number of known genomes is still immeasurably small compared to the number of species for which no sequence information is available. Genomic information is missing for less well studied species and incomplete for model-organisms such as the honeybee (*Apis mellifera*) (Beye *et al*, 1998; Peiren *et al*, 2005) and the African clawed frog (*Xenopus laevis)* (Liska *et al*, 2004a).

Since it cannot be expected that this lack of genomic sequence information will be resolved in the near future, approaches were developed to enable the identification of proteins by means of homology to known proteins. Search parameters that are used are amino acid composition, the pI, the mass, the hydrophobicity, the masses of peptides after digestion with endoproteases, and homology between observed sequences of the unknowns and proteins in databases. A theoretical study revealed that, of all these parameters, only the mass of a protein, its amino acid composition and its sequence are sufficiently conserved among homologous proteins to allow efficient cross-species protein identification (Wilkins *et al*, 1997).

In all approaches for cross-species protein identification the result finally depends on the availability of databases containing sequences of related species. If the organism being studied is only distantly related to organisms with a sequenced genome, the likelihood of protein identification decreases. Habermann *et al* evaluated the applicability of MS BLAST for the identification of proteins isolated from organisms that are only distantly related to database entries (Habermann *et al*, 2004). Statistical analysis revealed that, in the mammalian subkingdom, over 80% of the proteins can be identified by homology to available genomic information. A similar conclusion can be drawn for the entire vertebrate lineage. However, for earlier diverged lineages, such as fungi or bacteria, the potential to ascertain homology decreases. For the other programs used here, no such in depth analysis was performed. Cross-species protein identification by comparison of the position of spots on 2D-gels (Mathesius *et al*, 2002), or with PMF-analysis (Cordwell *et al*, 1995; Wasinger *et al*, 1995) is limited and can only be applied for very closely related species.

In 1998, Wilkins *et al* introduced MultiIdent (http://www.expasy.ch/sprot/multiident.html), a database search tool specifically designed for cross-species protein identifications. MultiIdent uses multiple protein parameters such as amino acid composition, peptide masses, sequence tags, estimated protein pI and mass (Wilkins *et al*, 1998). They demonstrated that amino acid composition is highly conserved among homologous proteins (Wilkins *et al*, 1997). Therefore, this information is used as a first parameter to create a short list of candidate proteins. The other parameters were subsequently used to discriminate between these candidate proteins. DeNovoID also uses the amino acid composition of peptides to perform database searches. However, it requires stringent manual validation of the results and was only used for the identification of human proteins (Halligan *et al*, 2005).

When a protein shares less than 70% identity to database entries it can only be identified by sequence-homology searches. The difficult task of *de novo* sequence analysis of

peptides after MS/MS (Part 3.1.1) and the optimization of commonly used database searching tools for non-error-tolerant searches necessitated the development of specialized software. Even programs designed for homology searches, BLAST or FASTA, have difficulties identifying related sequences sharing less than 90% identity when the query only consists of a single short amino acid sequence (Altschul *et al*, 1994). Software, specifically developed for cross-species protein identifications with MS/MS data, has to meet some requirements. Firstly, the software must be able to differentiate homologous amino acid substitutions from random changes. For this differentiation matrixes are used that represent the probability that one amino acid is changed for another during evolutionary times. The first matrix that was constructed is the PAM-matrix (Percent Accepted Mutations) composed by alignment of numerous closely related proteins. Because the PAM-matrix has problems coping with evolutionary divergent proteins, a new scoring matrix was described in 1992 which is based on the alignment of blocks of closely related proteins: BLOSUM (BLOcks SUbstitution Matrix) (Henikoff *et al*, 1992). Secondly, since identification of proteins with tandem mass spectrometry is based on the fragmentation of multiple peptides, the order of the sequences in the query is unknown; the software must be able to sort the peptide sequences in order to attain an optimum result. Finally, an appropriate search algorithm must be able to deal with the characteristics of MS determined sequences. In most approaches, isobaric amino acids can not be readily distinguished. The pairs Ile/Leu, Gln/Lys and Phe/oxidized Met have the same nominal mass and the possibility to consider them as equal should be encrypted in the software. In the studies reported here, three different programs for homology searching were used; FASTS, MS BLAST and MS-Homology. The characteristics of these programs are discussed in the Sections 3.2 and 3.3.

3.2. In-gel guanidination; development and proof of principle

De novo sequence analysis of N-terminal sulfonated peptides after in-gel guanidination

Kjell Sergeant, Bart Samyn, Griet Debyser and Jozef Van Beeumen

University of Gent
Department Biochemistry, Physiology and Microbiology
Laboratory of Protein Biochemistry and Protein Engineering
K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

Published in Proteomics (**2005**), 5(9), 2369-80.

Introduction

Proteomics is playing a pivotal role in the postgenome era in helping to define the functional role of genes. Contemporary proteomics is largely based upon separations using 2D-PAGE followed by mass spectrometric analysis for protein identification. Typically, proteins are digested using trypsin and the resultant peptides provide a characteristic 'mass fingerprint', which can be used to identify proteins. Notwithstanding the fact that the peptide mass fingerprint (PMF) approach is useful for identifying proteins in simple mixtures, the identification of proteins in more complex mixtures often requires partial peptide sequence data obtained by MS/MS methods. If accurate genome sequence information is available, database search algorithms can be used in conjunction with MS/MS for readily identification. Sophisticated algorithms (e.g. SEQUEST, Mascot) have been developed for identifying proteins from peptide MS/MS data; whereby peptides are identified by correlating the uninterpreted MS/MS spectra with simulated (predicted) product ion spectra derived from peptides of the same mass contained in the available databases. For proteins not contained within sequence databases, it is necessary to determine partial or complete amino acid sequences using either manual or automated *de novo* peptide sequence analysis methods. While the above-mentioned algorithms for protein identification from peptide MS/MS data have enjoyed considerable success, their utility is directly related to the quality of the product ion spectra. Thus, if product ion spectra are formed that are not readily interpretable, low or insignificant search scores can result.

Attempts to simplify and improve the fragmentation pattern have largely concentrated on the introduction of charged groups at the N- or C-terminus of the peptide. Recently developed derivatization strategies involve methods to improve the ionization efficiency (guanidination) (Beardsley *et al*, 2000; Brancia *et al*, 2000; Hale *et al*, 2000), to enhance MS/MS fragmentation (CAF) (Keough *et al*, 1999; Hellman *et al*, 2002; Keough *et al*, 2003), and methods that allow relative quantification (ICAT) (Gygi *et al*, 1999). General requirements for derivatization reagents include the following: (1) the reaction with target functional groups must be fast and reproducible, with no side reactions, and with yields as close as possible to 100%; (2) the derivatives must be stable under MS and MS/MS conditions, as well as through multiple separation and sample cleanup steps; (3) the modification may not impair chromatographic separation; and (4) the tag should enhance signal intensity in the MS mode and improve the quality of the MS/MS spectra compared to the corresponding unlabeled peptides (Pashkova *et al*, 2004).

Keough *et al* developed a derivatization strategy that facilitates *de novo* sequencing of singly charged tryptic peptides. Introduction of a strong acid group at the N-terminus of tryptic peptides was shown to facilitate protonation of backbone amide bonds. This protonation destabilizes amide bonds, leading to extensive structure-specific fragmentation under post-source decay MALDI and electrospray tandem mass spectrometry conditions (Keough *et al*, 1999; Bauer *et al*, 2000). In 1999, Keough *et al*. reported the use of chlorosulfonylacetyl chloride (CSAC) as a N-terminal derivatization reagent. CSAC is a commercially available reagent that allows direct sulfonation of subpicomole quantities of tryptic peptides. Since sulfonation reagents react with amino groups this derivatization results in modification of both the N-termini and the α-amines of lysine-containing peptides. Therefore, Keough *et al*. expanded their approach and combined guanidination of lysine residues with the addition of a sulfonic acid group to the N-terminus (Keough *et al*, 2000b). Following guanidination of lysine α-amines, introduction of sulfonic acid groups to tryptic peptides is possible solely at the N-terminus. Beardsley and Reilly recently described an

optimized procedure in which selective and quantitative (e.g. complete) guanidination is performed at high pH with a concentrated *O*-methylisourea solution (Beardsley *et al*, 2002). This treatment efficiently transforms lysine into homoarginine, which is 42 amu heavier than lysine, but does not affect the peptide amino terminus or other side groups. A major drawback is that the reagents used for guanidination results in significant contamination of the peptides, thus necessitating a purification step prior to MALDI analysis (Cagney *et al*, 2002; Hellman *et al*, 2002; Keough *et al*, 2003). Although the benefits of sample cleanup and enrichment with reversed-phase microextraction columns have been demonstrated, analyte losses are often unavoidable. The group of Reilly observed that high-mass peptide signals do not increase and are in some cases even absent following guanidination (Beardsley *et al*, 2000).

In a recent report we have demonstrated that full-length *de novo* sequencing is possible using MALDI-TOF/TOF analysis. The interpretation of MS/MS data is facilitated by N-terminal sulfonation after protection of lysine side chains (Samyn *et al*, 2004). Here, we report an improved method in which the Lys side chains of gel-separated proteins are guanidinated in-gel prior to trypic digestion. The protocol was first optimized using a number of SDS-PAGE separated test proteins. The first objective of this study was to examine whether the guanidination reaction can be performed in-gel, on the intact protein, in order to eliminate the time- and sample-consuming desalting step that is necessary when the reaction is performed on tryptic peptides. The influence of the guanidination and sulfonation modifications on the PMF composition is compared with respect to the normal PMF approach. As a proof of principle, the improved approach is applied to identify a number of 2D-PAGE separated proteins from *Shewanella oneidensis*. For protein identification we used a recently described algorithm, FASTS, which can search a database with all MS-derived *de novo* peptide sequences simultaneously (Mackey *et al*, 2002). Finally, we also demonstrate the possibility to characterize post-translational modifications (PTMs) using this approach.

Materials and methods

Test protein mixture

A set of standard proteins was prepared containing the following proteins (Sigma, Bornem, Belgium): horse cytochrome c (HCC)(96% pure), horse heart myoglobin (MYO)(90% pure), bovine β-casein (90% pure), bovine serum albumin (BSA)(98% pure) and yeast alcoholdehydrogenase (YADH)(90% pure). For each protein, a stock solution of 50 pmol/µl was prepared. The amount of sample applied on the gel was 50 pmol of each protein.

SDS-PAGE

A mixture of test proteins was electrophoretically separated according to Laemmli. 12% Tris-glycine gels of a thickness of 1 mm containing 10 wells were casted. Electrophoresis was carried out using a SE250 Mighty Small II apparatus (Hoefer Scientific, San Francisco, CA) at room temperature. Protein samples were mixed 1:1 (v/v) with sample buffer containing β-mercaptoethanol as the reducing agent and bromophenol blue to visualize the electrophoresis front. The sample was heated briefly (90-95°C, 5 min) before it was loaded on the gel. The electrophoresis running buffer was 25 mM Tris base, 192 mM glycine, and 0.1% SDS (w/v). Electrophoresis was carried out at 150 V for +/- 1.5 hours, until the dye marker had reached the edge of the gel. After fixation (2% H_3PO_4/50% ethanol/MQ; 30 minutes), proteins were stained with CBB G-250 at 0.2% (w/v) in 34% methanol/17%

ammonium sulfate, containing 3% fosforic acid, for +/- 30 minutes. Destaining was carried out overnight with a 30% methanol solution.

Guanidination in solution

The separated proteins were excised and the gel pieces washed twice with 150 µl 200 mM NH_4HCO_3/50% ACN for 30 minutes at 30°C. Subsequently, the gel pieces were dried in a Speedvac (Thermo Savant, Holbrook, NY). The proteins were digested in-gel with trypsin as described below and the peptides extracted twice with 35 µl 60% ACN/0.1% DIEA. The extracts were pooled, dried in the SpeedVac and dissolved in 5 µl MQ. A 7.5 M *O*-methylisourea hemisulfate stock solution was prepared freshly by dissolving 0.050 g *O*-methylisourea hemisulfate (Across, Geel, Belgium) in 51 µl water. Guanidination was performed by adding 5.5 µl 7N NH_4OH and 1.5 µl of the *O*-methylisourea solution. After a short incubation of 15 min at 65°C, the samples were desalted using C18 micro purification tips (ZipTip, Millipore).

In-gel guanidination

Guanidination was performed by adding 5 µl MQ, 11 µl 7N ammonium hydroxide (Merck, Darmstadt, Germany) and 3 µl of the 7.5 M *O*-methylisourea hemisulfate solution to the excised spots. The samples were vortexed briefly and incubated at 65°C. After an incubation of two hours the guanidinated samples were taken from the oven and the remainder of the reaction mixture was discarded. After guanidination, the gel pieces were desalted and destained in one step. The gel spots were washed twice with 150 µl 200 mM ammonium bicarbonate in 50% ACN/water (30 min at 30°C) and dried in the SpeedVac.

Trypsin digestion and sulfonation

To the dried gel spots 8 µl digestion buffer (50 mM ammonium bicarbonate, pH 7.8) containing 150 ng modified trypsin per µl (Promega, Madison, WI, USA) was added and the tubes were kept on ice for 45 minutes to allow the gel pieces to be completely soaked with trypsin. Digestion was performed overnight at 37°C, the supernatans was recovered and the resulting peptides extracted twice with 35 µl 60% ACN/0.1% DIEA. The extracts were pooled and dried in the SpeedVac. The peptides were redissolved in 4 µl 12.5 mM ammonium bicarbonate 50% ACN and 2 µl was mixed with 2 µl of the sulfonation solution. The sulfonation reagent was prepared by dissolving 2 mg 2-sulfobenzoic acid cyclic anhydride in 1 ml dry tetrahydrofuran (THF) to attain a 0.01 mM solution. The tubes were briefly vortexed, and reacted for 5 minutes at room temperature.

Bacterial growth, preparation of extracts and 2D-PAGE gel electrophoresis

Bacterial growth and preparation of the protein extract was performed as follows. *S. oneidensis* MR-1 was grown aerobically overnight in 20 ml Luria-Bertani (LB) medium in a rotary shaker at a speed of 200 rpm at 28°C until exponential phase (OD_{600} = ±1). The cells were centrifuged and washed twice using a 50 mM Tris-HCl solution (pH 8) containing 5 mM EDTA. The bacteria were lysed using 9M urea, containing 2% CHAPS, 1% DTT and 0.8% ampholines, after which the pellets were centrifuged at 14000 rpm. A volume of 20 µl of the bacterial extract (± 500 µg of protein as determined by a Bradford test) was loaded on a 18 cm IPG strip, pH range 4-7 (Pharmacia, Uppsala, Sweden). After resolubilization in a rehydration solution (6 M urea, 4% (w/v) CHAPS, 75 mM DTT), the mixture was applied to an IPG strip

for 6 to 8 hours at room temperature. IEF was performed using a standard program as provided by the manufacturer at 18°C. The equipment for the rehydration and for running the IPG gels (Multiphor II) was purchased from Amersham Biosciences (Uppsala, Sweden). After focusing, the strips were equilibrated in 5 ml of 50 mM Tris-HCl (pH 8.8)/6 M urea/30% glycerol/2% SDS and 1% DTT. After 10 minutes the reducing solution was replaced by the acylating solution (same solution with DTT being replaced by 2.5% IAA) for 10 minutes at room temperature. For the second dimension, 12.5% SDS-PAGE gels (18 cm x 18 cm) were run in a Protean II (Bio-Rad, Nazareth, Belgium) electrophoresis apparatus at 8°C and +/- 20 mA/gel, until the bromophenol blue front reached the bottom of the gel. The gels were fixated for one hour, stained with CBB overnight and destained for 4 hours as described above.

Mass spectrometry

The Applied Biosystems 4700 Proteomics Analyzer with TOF/TOF optics was used in this study for MALDI MS and MS/MS applications (Applied Biosystems, Foster City, CA). This mass spectrometer uses a 200-Hz frequency tripled Nd:YAG laser operating at a wavelength of 355 nm. For MS/MS, ions generated by the MALDI process were accelerated at 8 kV through a grid at 6.7 kV into a short, linear, field-free drift region. In this region, the ions passed through a timed-ion-selector device that is able to select one precursor for subsequent fragmentation in the collision cell. After a peptide at a given m/z was selected it passed through a retarding lens where the ions were decelerated and then passed into the collision cell, which was operated at 7 kV. The collision energy is defined by the potential difference between the source and the collision cell (1 kV). After passing through the collision cell, the ions (both intact peptide and fragments) were accelerated in the second source region at 15 kV, passed through a second, field-free, linear drift region, into the reflector, and finally, to the detector. The detector amplifies and converts the signal to electric current, which is observed and manipulated on a PC-based operating system. For high resolution MS analysis, the instrument was operated in reflectron mode. After the MALDI process generates the peptide ions, the latter are accelerated at 20 kV through a grid at 14 kV into the first, short, linear, field-free drift region. After this point, the rest of the instrument can be treated as a continuation of this region until the ions enter the reflector and then reach the detector where, as before, the signal at the detector is amplified and converted to electrical current.

Samples were prepared by mixing 0.7 µl of the sample with 0.7 µl matrix solution (7 mg/ml CHCA in 50% ACN containing 0.1% TFA) and spotted on a stainless steel 192-well target plate. They were allowed to air-dry at room temperature and then inserted in the mass spectrometer and subjected to MALDI MS analysis. Prior to analysis, the mass spectrometer was externally calibrated with a mixture of Angiotensin I, Glu-fibrino-peptide B, ACTH (1-17), and ACTH (18-39). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrino-peptide.

Database analysis

After manual interpretation, the multiple *de novo* sequences from one spot were submitted in a single database search using the FASTS-program (http://fasta.bioch.virginia.edu/fasta_www/cgi/) or the program MS BLAST (http://dove.embl-heidelberg.de/Blast2/msblast.html). Both programs were used with standard settings to search against the NCBI non-redundant protein database. Identification by peptide mass fingerprint analysis was

performed using a local Mascot server against a comprehensive *Shewanella* sequence database downloaded from the Institute of Genomic Research (http://www.tigr.org). Cysteine carboxyamidation and oxidation of methionines were allowed as variable modifications and a mass tolerance threshold of 100 ppm was used.

Results and discussion

Guanidination in-gel versus modification in solution

To examine the role of performing the guanidination in-gel, prior to the tryptic digest, a number of test proteins were separated by SDS-PAGE and subjected to both protocols. The tryptic digest was also performed in solution on 50 pmol of all model proteins. The results of these experiments are summarized in Table 3.4. All of the sulfonated peptides were subjected to MS/MS using a MALDI-TOF/TOF instrument. Fragmentation of the derivatized peptides occurred under metastable decay conditions and a frame collision energy of 1 keV (no gas in the collision cell). In all experiments CHCA was used as matrix, as this is the most commonly used matrix for peptide analysis. A schematic representation of the different approaches is given in Figure 3.14.

Table 3.4. Standard proteins

Protein[a]	M_r (kDa)	trypsin in-solution guanidin./sulf. sol.		trypsin in-gel guanidin./sulf. sol.		guanidin. in-gel sulf. in solution	
		# AA	% *de novo*	# AA	% *de novo*	# AA	% *de novo*
HCC	12.3	45	43.27	57	54.81	29	27.88
MYO	16.9	22	14.38	72	47.06	44	28.76
α-cas	24.0	6	2.68	18	8.04	33	14.73
YADH	37.1	60	17.29	118	34.01	84	24.21
BSA	68.2	67	11.04	93	15.32	79	13.01

[a] The used abbreviations are defined in the Materials and methods section

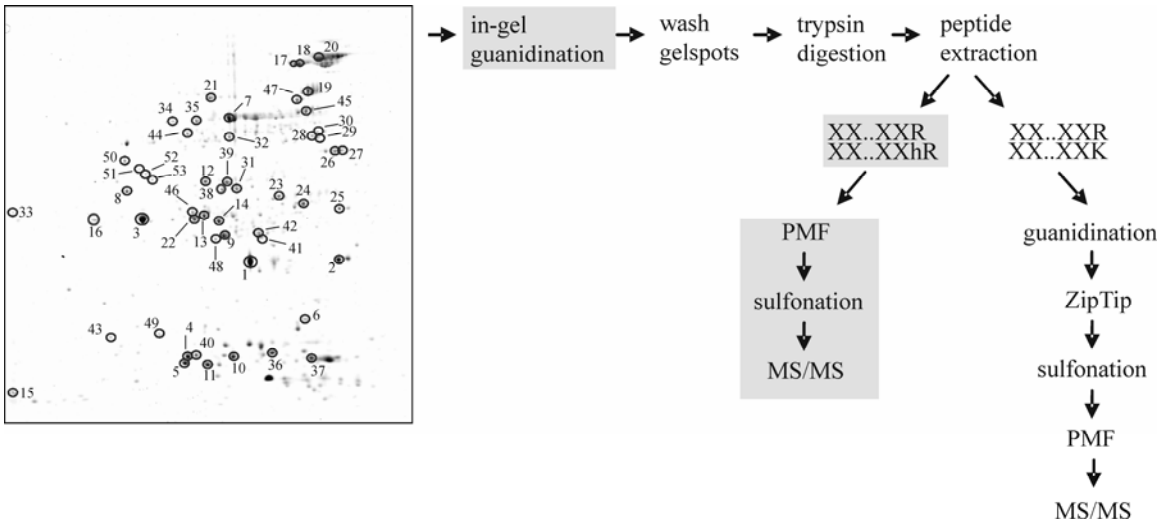


Figure 3.14. Schematic representation of the different steps in the derivatization stategies. Steps specific for the in-gel guanidination protocol are marked in grey boxes. The standard protocol, involving trypsin digestion followed by guanidination, desalting (ZipTip) and sulfonation, is indicated in the other pathway. Analyzed spots from the *Shewanella oneidensis* in the 2D-gel are numbered. X: any amino acid, hR: homoarginine.

In all approaches, a contiguous series of y-ions was observed in the fragment spectra after guanidination and sulfonation. The y-ion series could easily be manually interpreted, facilitating *de novo* sequencing. In all fragment spectra an initial loss of the sulfonic acid derivative was observed (Δm = 184 Da). As observed before, in some experiments the sulfonic acid-derivatized peptides had poorer positive-ion sensitivity than the corresponding native peptides. After sulfonation, some of the tryptic peptides were no longer observed in positive mode reflectron analysis. These fragments could be detected as their deprotonated ions when the analysis was performed in the negative mode. However, selection of the corresponding protonated precursor ion for MS/MS analysis (positive mode) results in the formation of a complete series of y fragment-ions. Figure 3.15 shows the derivatized tryptic fragments of BSA. MS analysis in negative reflectron mode indicates the presence of two more peptides with respect to the spectrum in positive reflectron analysis (Figure 3.15a & b) (Δm = 2 Da). Figure 3.15c shows the fragmentation spectrum (in positive mode) of the theoretical precursor at m/z 1751.6 from which the complete sequence could be deduced. The same result was obtained by MS/MS analysis of the 1663.66 precursor (results not shown). Most likely the protonated precursor is not detected due to metastable decomposition, but does yield excellent MS/MS spectra in the positive ion mode (Samyn *et al*, 2004).

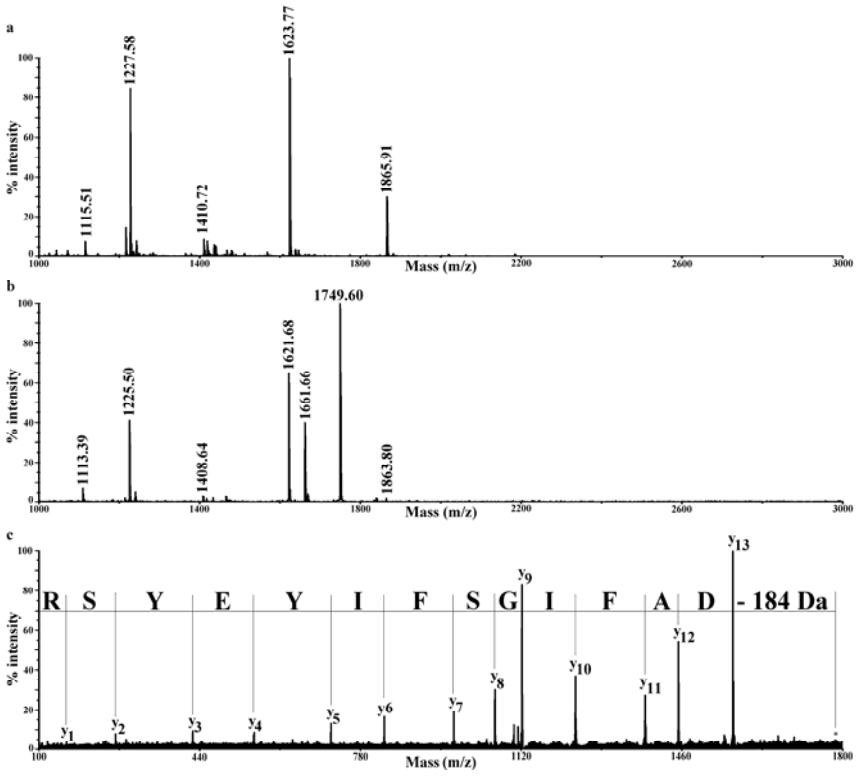


Figure 3.15. a) MALDI reflectron MS spectrum (positive ion mode) of BSA after performing the in-gel procedure, b) MALDI MS reflectron spectrum (negative ion mode) of the same sample. Peaks at m/z 1749.60 and m/z 1661.66 correspond to sulfonated peptides that were not detected in positive ion mode c) MS/MS spectrum (positive ion mode) of the precursor at m/z 1751.60 (1749.60 + 2), the loss of the sulfonation label (-184 Da) is indicated and y-ions are labeled.

Performing the tryptic digest in-gel rather than in solution resulted in a higher number of peptides and subsequently a higher sequence coverage (Table 3.4). When the guanidination reaction was performed in-gel prior to the enzymatic cleavage, all Lys residues are converted to homo-Arg. As observed before, homo-Arg derivatives were not or less efficiently cleaved

by trypsin (Hara *et al*, 1995). The resulting peptide fragments, containing an internal homo-Arg, had a higher molecular weight than the tryptic fragments containing no 'missed-cleavages'. Mostly, this resulted in a decrease of the sequence coverage except in the case of α-casein. Tryptic digestion of underivatized α-casein resulted in the formation of a large number of small peptides, with a molecular mass below that of the matrix interference clusters (+/- 1 kDa). MS/MS analysis of derivatized peptides with an internal homo-Arg produced complete y-ion series. However, in most spectra the y-ion series were often accompanied with (y_i –17)-ion series due to the neutral loss of NH_3. Although this (y_i –17) series shows up in the fragment spectra until the basic residue is cleaved off, its presence does not interfere with the sequence interpretation. The tryptic peptide DTHKSEIAHR of BSA has an internal homo-Arg at position 4. As illustrated in Figure 3.16, fragmentation of the derivatized peptide (precursor m/z 1419.59) yielded the complete y-ion series along with a more pronounced (y-17)-ion series, up to the (y_7-17)-ion, the place where the homo-Arg was fragmented. As recently reported, the neutral loss (Δm = 17 Da) might prevent a correct sequence interpretation for peptides having a Asn-Pro bond in their sequence (Δm Asn/Pro = 17 Da) (Samyn *et al*, 2004). However, the occurrence of 'non-sequence' specific ion fragments, such as the neutral loss of ammonia from peptides with internal Arg or homo-Arg, can be used to improve predictive models of peptide fragmentation for *de novo* sequence analysis. The conversion of Lys to homo-Arg in-gel was estimated to be essentially complete, as no signals from unguanidinated lysine containing peptides were observed in any experiment (results not shown).

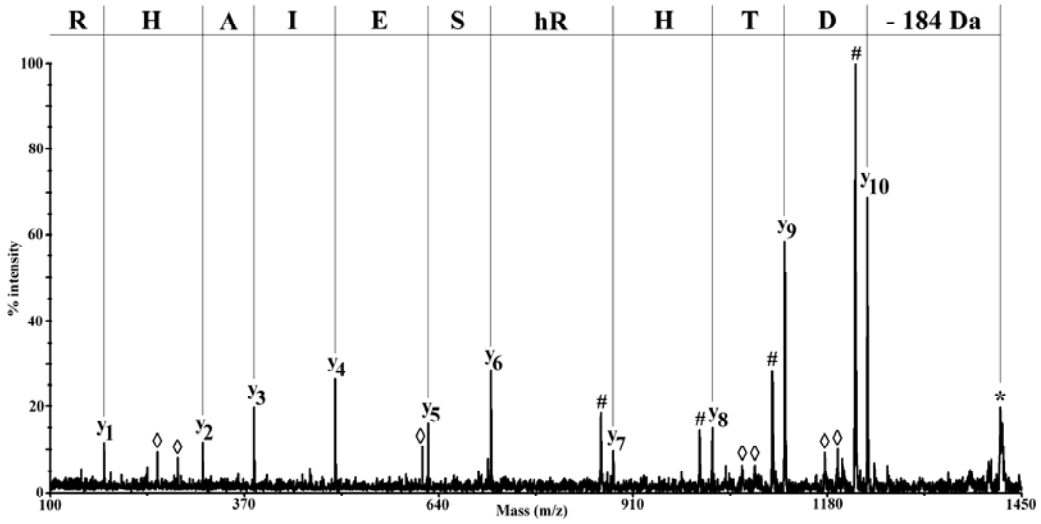


Figure 3.16. MS/MS spectrum (positive ion mode) of the sulfonated peptide DTHhRSEIAHR (BSA, D25 – R34) (hR, homoarginine). The sulfonated precursor (m/z 1419.59) is labeled with *. (y-17)-ions resulting from the neutral loss of NH_3 are indicated as #. Where appropriate, internal ions are labeled with ◊.

Shewanella oneidensis MR-1

The improved guanidination procedure was also applied to study 2D-PAGE separated bacterial proteins from *Shewanella oneidensis* MR-1, a bacterium whose genome (4.9 Mbp, 5177 ORFs) has recently been sequenced by the Institute for Genomic Research (http://www.tigr.com). Since the microorganism is able to reduce a variety of metal substrates, it is of considerable interest to researchers involved in bioremediation (Glasauer *et al*, 2002). Although a limited number of proteomic studies of *Shewanella* have been reported (VerBerkmoes *et al*, 2002; Mohan *et al*, 2003; Vanrobaeys *et al*, 2003), only a fraction of the *Shewanella* genes have been characterized at the protein level so far.

Table 3.5. 2D-PAGE separated proteins of *Shewanella oneidensis*

spot[a]	protein[b]	PMF[c]	identification FASTS[d]	E-score[e]	#peptides[f]	# AA[g]
1	15640750	+	antioxidant, AhpC/Tsa family	3.70e-39	6	63
2	24373389	+	conserved hypothetical protein	1.80e-26	4	38
3	24376191	+	conserved hypothetical protein	1.50e-18	3	32
4	24373827	+	nucleoside diphosphate kinase	3.10e-40	4	61
5	24374659	+	DNA-binding prot, H-NS family	6.30e-22	4	35
6	24372148	-	hypothetical protein SO0554	8.30e-06	2	15
7	24371815	+	translation elongation factor Tu	7.70e-27	6	69
8	24372773	+	conserved hypothetical protein	5.10e-20	4	33
9	24372802	+	purine nucleoside phosphorylase	7.00e-09	2	22
10	24375179	+	universal stress protein family	5.90e-26	2	37
11	24373389	+	conserved hypothetical protein	1.20e-26	4	51
12	24372359	+	malate dehydrogenase	1.80e-38	4	50
13	24371943	+	methylisocitrate lyase	1.80e-24	4	43
14	24373372	+	phage shock protein A	4.80e-31	5	50
15	24374936	-	carbon storage regulator	4.40e-12	3	28
16	24376191	+	conserved hypothetical protein	3.00e-19	3	28
17	24373949	+	ribosomal protein S1	7.20e-11	4	32
18	24373949	+	ribosomal protein S1	4.00e-37	8	82
19	24372295	+	chaperonin GroEL	1.40e-51	5	68
20	24372709	+	chaperone protein DnaK	1.80e-47	5	69
21	24376221	+	ATP synthase F1, alpha subunit	6.70e-37	7	59
22	24376191	+	conserved hypothetical protein	6.10e-16	4	33
23	24374657	+	electr transf flavoprot, α subunit	5.30e-22	3	36
24a	24373875	+	translation elongation factor P	6.70e-11	4	27
24b	24372295	-	chaperonin GroEL	5.30e-06	1	17
25	24373389	+	conserved hypothetical protein	7.90e-27	4	37
26	24375049	+	OmpA family protein	4.60e-07	4	25
27	24375049	+	OmpA family protein	2.70e-16	3	29
28	24371854	+	DNA-RNA polymerase α-chain	2.40e-43	6	58
29	24371854	+	DNA-RNA polymerase α-chain	3.10e-10	3	21
30	24371854	-	DNA-RNA polymerase α-chain	2.30e-27	4	40
31	24373198	-	translation elongation factor Ts	2.40e-35	5	48
32	24373496	+	succinyl-CoA synth, beta-subunit	2.00e-36	7	56
33	24373197	-	ribosomal protein S2	6.90e-14	4	36
34	24375430	-	serine prot, HtrA/DegQ/DegS fam	6.90e-35	6	57
35	24375430	-	serine prot, HtrA/DegQ/DegS fam	3.00e-43	5	58
36	24372294	+	chaperonin GroES	8.80e-25	7	64
37	24371821	-	ribosomal protein L7/L12	2.20e-16	3	32
38	24373198	+	translation elongation factor Ts	4.80e-21	3	31
39	24372359	+	malate dehydrogenase	2.10e-39	5	50
40	24375415	+	ribosomal protein L9	7.10e-55	7	76
41	24372202	+	stringent starvation protein a	7.00e-20	4	43
42	24374593	+	molybd ABC transp, periplasmic	7.00e-28	5	57
43	24374881	+	conserved hypothetical protein	2.30e-12	3	26
44	24374593	-	3-oxoacyl-(acyl-carrier-prot) synth	9.80e-21	6	44
45a	24376219	+	ATP synthase F1, beta subunit	2.80e-17	6	39
45b	24375475	+	aerobic respirat control prot ArcA	7.20e-53	6	67
46	24376191	+	conserved hypothetical protein	2.10e-14	3	28
47	24373359	+	trigger factor	4.50e-37	5	55
48	24374298	+	uracil phosphoribosyltransferase	5.40e-16	3	27
49	24374881	+	conserved hypothetical protein	1.00e-09	2	20
50	24372520	+	fruct-bisphosphate aldol, class II	1.00e-14	4	30
51	24373111	+	isocitrate dehydrogenase	1.10e-23	3	35
52	24372333	+	iron(III) ABC transp, periplasmic	7.10e-18	3	29
53	24373497	+	succinyl-CoA synth, alpha-subunit	2.40e-12	3	28

[a] spot number according to the position on the 2D-PAGE (Figure 3.14)
[b] NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/)
[c] + (-) indicates a positive (negative) MASCOT identification
[d] protein with lowest E-value in FASTS search result
[e] In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI non-redundant protein databases N ≈ 2075116.
[f] number of peptide sequences used in the query

[g] total number of amino acids used in the query

The total protein extract from aerobically grown *Shewanella* MR-1 was separated by 2D-PAGE (Figure 3.14). Two identical extracts were analyzed; the first gel was used for identification using the 'standard' tryptic mass fingerprint approach (no derivatization). After Coomassie staining, 53 spots and two blanks were randomly selected. In the 53 spots, 43 proteins with a M_r ranging from 7.1 to 68.8 kDa, could be identified by performing a database search on a local Mascot server (Perkins *et al*, 1999). Using this approach, the protein in spot 19, e.g., was identified as chaperonin GroEL (Table 3.5). The same spots in the second gel were subjected to the new protocol and a summary of the results is presented in Table 3.5. By way of an example, Figure 3.17a shows the derivatized peptides from spot 19 after in-gel guanidination and sulfonation. The sulfonated tryptic peptides were subjected to MS/MS analysis. Figure 3.17b shows the fragmentation spectrum of the precursor at m/z 2712.44. Since the peptide contained an internal homo-Arg, the (y_i-17)-ions were predominant in the spectrum whereas the y-ions were not or hardly observed until the homo-Arg was fragmented. Therefore, the loss of the sulfonation derivative (-184 Da) and the neutral loss of ammonia (-17 Da) are summed together as – 201 Da (Figure 3.17b).

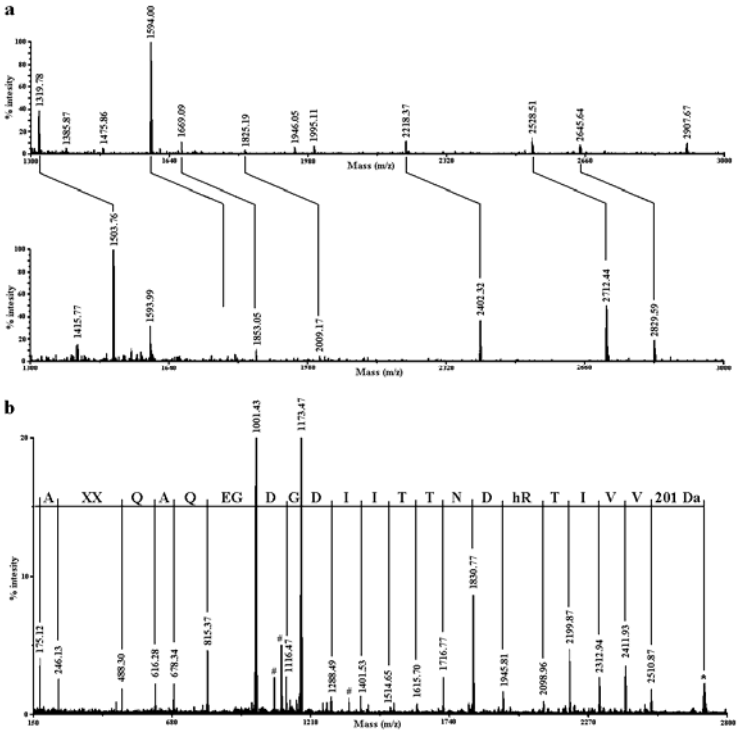


Figure 3.17. a) MALDI reflectron MS spectra (positive ion mode) of the in-gel guanidinated (upper panel) and sulfonated (lower panel) tryptic peptide mixture of spot 19 indicated in Figure 3.14. Increases of 184 Da, the mass of the sulfonating group, are indicated. b) MS/MS spectrum (positive ion mode) of the sulfonated precursor (m/z 2712.44) indicated with *. y-ions are indicated, except for masses higher than 1945.81 where the labeled peaks are the (y-17)-ions. Internal ions are indicated as #. hR: homoarginine, X: any amino acid.

The *de novo* derived sequences were used as a query in the FASTS algorithm (Mackey *et al*, 2002). FASTS is a recently reported sequence similarity search algorithm which searches databases using peptide sequences of unknown order, evaluating all possible arrangements of the peptides. FASTS is designed to use *de novo* sequence data from organisms lacking comprehensive proteome sequence data. The algorithm uses the heuristic FASTA comparison strategy to accelerate the search but uses alignment probability, rather than a similarity score, as the criterion for alignment optimality. Because the true order of the query peptides used by FASTS is not known, FASTS only requires that the aligned peptides

do not overlap. The guanidination allows to discriminate Lys from Gln but, unfortunately, the isobaric amino acids Leu and Ile cannot be differentiated from the MS/MS spectrum. For identifications with FASTS, Ile was chosen for every 113 Da difference observed in the fragment spectra. Gaps in the y-ion series were indicated with XX as indicated in Figure 3.17b.

Using this approach, we were able to identify all proteins in the 53 spots. In Table 3.5, the number of fragmented peptides and the number of determined amino acids (the sequence length) for each spot is indicated. For spot 19, for example, the four peptides with highest intensity were subjected to MS/MS analysis and yielded 60 amino acids *de novo*. Upon sulfonation, the guanidinated peptide at m/z 1594.00 was no longer observed in positive mode reflectron analysis (Figure 3.17a). In negative reflectron mode, the sulfonated peptide was observed at m/z 1775.86. Selection of the theoretical 1777.86 Da precursor in MS/MS analysis yielded an additional 8 AA sequence for the FASTS query (all data on the peptides used for MS/MS and the *de novo* determined sequences are summarized in Appendix 1). The FASTS query, made against the National Center for Biotechnology Information (NCBI) non-redundant protein database, identified the spot as the chaperonin GroEL with an E-score of 1.4e-51 (Table 3.6). This is an identification with a twofold higher confidence level than the second best hit, GroEL from *Pseudoalteromonas* sp. (E-score = 3.3e-24). In total, 644 proteins, almost all of them 60 kDa heat shock proteins, had an E-score above the treshold value of 1.0e-5. This result indicates that this approach should allow the identification of proteins from organisms whose genomes have not been sequenced. FASTS requires only 15-20 total residues in three to four peptides to robustly identify homologues sharing 50% or greater protein sequence identity. Searches with a smaller number of longer peptides are more sensitive, particularly at greater evolutionary distance. For the identification of the Shewanella proteins, two to eight peptides with an average length of ten amino acids were used in the queries. For all proteins, the lowest E-score corresponded with a protein from Shewanella oneidensis and confirmed the identifications made by Mascot (Table 3.5).

Table 3.6. FASTS search result for spot 19

Protein[a]	identification	organism	E-score[b]
24372295	chaperonin GroEL	*Shewanella oneidensis* MR-1	1.4e-51
13366173	GroEL	*Pseudoalteromonas sp.* PS1M3	3.3e-24
15213863	GroEL	*Sodalis glossinidius*	4.3e-24
20137922	60 kDa chaperonin	*Sodalis glossinidius*	4.7e-24
3913235	60 kDa chaperonin	*Actinobacillus pleuropneumoniae*	6.1e-24
32033972	Chaperonin GroEL	*Actinobacillus pleuropneumoniae serovar 1*	6.1e-24
31339380	chaperonin hsp60	*Colwellia maris*	1.1e-23
5524758	GroEL protein	*Pseudoalteromonas haloplanktis*	1.6e-23
38349494	GroEL	*secondary endosymbiont of Bemisia tabaci*	5.4e-23
7443844	chaperonin GroEL-like prot	*Sitophilus oryzae*	1.0e-22
2564288	Hsp60 protein	*Pseudomonas stutzeri*	4.2e-22
12721449	GroEL	*Pasteurella multocida subsp. multocida str.* Pm70	6.5e-22
477903	heat shock protein GroEL	*Haemophilus ducreyi*	6.5e-22
1144302	GroEL	*Pasteurella multocida*	6.5e-22
33152794	GroEL protein	*Haemophilus ducreyi* 35000HP	6.5e-22
52424514	GroEL protein	*Mannheimia succiniciproducens* MBEL55E	6.5e-22
34588127	60 kDa chaperonin	*Haemophilus ducreyi*	6.5e-22
29144631	GroEL protein	*Salmonella enterica subsp. enterica*	2.5e-21
29725663	heat shock protein GroEL	*Vibrio harveyi*	2.5e-21
15599581	GroEL protein	*Pseudomonas aeruginosa* PA01	2.5e-21

[a] NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/)
[b] In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI non-redundant protein databases N ≈ 2075116.

Figure 3.18 shows the PMF of the derivatized tryptic peptides from spot 37 in negative mode reflectron analysis. Fragmentation of three theoretical precursors (+ 2 Da) in positive mode and FASTS analysis with the *de novo* sequences indicated that this spot contained the ribosomal protein L7/L12 from *Shewanella oneidensis*. The MS/MS spectra from the cluster at m/z 1284.26, 1298.27 and 1312.26, each differing from one another by 14 Da, all yielded the same sequence, GATGIGIXEAhR (Figure 3.18b). The best hits in the FASTS results suggested that a Lys was present at position X in the sequence. The theoretical mass of the guanidinated, sulfonated and singly protonated peptide with Lys in position X is 1270.61 Da. The mass difference of 14 Da indicates the presence of a methylation on the internal Lys, preventing the guanidination of this residue. The other peaks in the cluster, 1298.27 and 1312.26, correspond to the peptide with a bi- and tri-methylated internal Lys respectively. This Lys (Lys82 according to gi:24371821 numbering) is a highly conserved residue amongst most bacterial species and in chloroplasts. Methylation of conversed Lys in ribosomal proteins has often been described and is implied as being essential for protein synthesis (Bocharov *et al*, 2004). The central roll of the ribosomal L7/L12 protein, a molecular switch in the binding of elongation factors, further suggests that this methylation has an important physiological function. Interestingly, a fourth peak was observed in the cluster with a M_r 42 Da higher than the methylated peptide (m/z 1326.27) suggesting that a minor fraction of the peptide was guanidinated. MS/MS analyis of this fragment indicated that a neutral loss of ammonia occurred (y_i-17 ion series) indicating the presence of an internal homo-Arg. However, an additional loss of 42 Da was observed in the fragmentation spectrum (Figure 3.18c) indicating that the guanidination on the mono-methylated peptide was not stable.

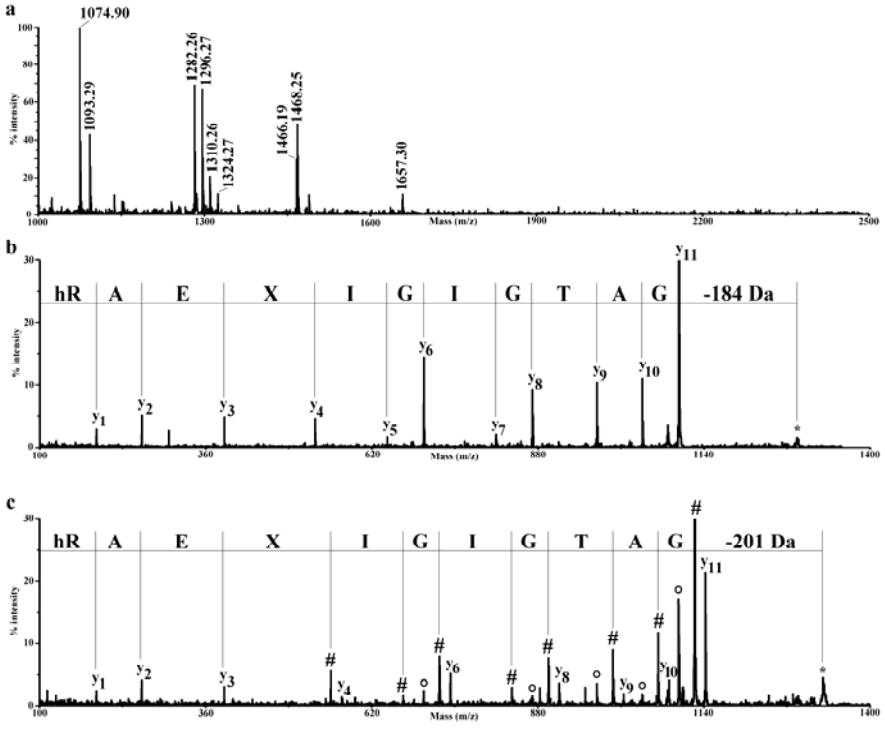


Figure 3.18. a) MALDI MS spectrum (negative ion mode) of the sulfonated tryptic peptide mixture of ribosomal protein L7/L12 (spot 37, Figure 3.14). b) MS/MS spectrum (positive ion mode) of the sulfonated precursor at m/z 1284.26, peptide Gly76–Lys86 monomethylated at Lys82. c) MS/MS spectrum of the sulfonated precursor at m/z 1326.27, peptide Gly76–Lys86, monomethylated and guanidinated at Lys82. Precursor ions are indicated with *, # designates (y-17)-ions and (y-42)-ions are labeled with an open circle.

Concluding remarks

Despite the benefits that guanidination conveys, one remaining problem is that the reagents used for guanidination result in significant contamination of the protein sample, thus necessitating a purification step prior to MALDI analysis. Beardsley & Reilly reported the use of a reversed-phase microextraction column for sample cleanup and enrichment after guanidination, however, they also reported that the intensities of nearly all peptides were diminished when using this approach at a high sensitivity level (Beardsley *et al*, 2002). In this paper, we have presented improvements to the guanidination technique. The new procedure is directly applied on gel separated proteins and therefore, the time and sample consuming desalting step can be omitted. In this way, removal of the molar excess of guanidination reagents can simply be accomplished during the destaining step of the gel spots. Although the number of recovered peptides is usually lower than by using the standard PMF approach (Table 3.4), sufficient sequence information was obtained for unambiguous identification of all 2D-PAGE separated proteins from *Shewanella oneidensis* (Table 3.5). Furthermore, we observed no underivatized lysines in any of the proteins, suggesting that the in-gel derivatization approach is quantitative. During guanidination, Lys is converted to homo-Arg and therefore can easily be differentiated from Gln. Unfortunately, as in other MS/MS approaches, Leu and Ile cannot be distinguished.

Although *de novo* sequencing of underivatized peptides using MALDI TOF/TOF has recently been demonstrated by Yergey *et al* (Yergey *et al*, 2002), the interpretation of fragment spectra from peptides originating from unknown proteins strongly depends on the use of automated search routines or on manual interpretation. TOF/TOF fragmentation analysis of underivatized peptides yields multiple, incomplete fragment ion series, which are often difficult to interpret. The use of high-energy CID (gas in the collision cell) results in the formation of high-energy fragment ions (w-ions), which may be indicative for the presence of Leu and Ile. However, it has been demonstrated that this type of CID also results in a decrease of the y-ion abundance at both higher gas pressures and higher collision energies. Therefore, in order to obtain longer, uninterrupted y-ion series, suitable for database similarity searching, all MS/MS collision experiments were performed in the low-energy CID mode, as previously reported (Samyn *et al*, 2004). The introduction of a sulfo group facilitates the MS/MS fragmentation of singly charged peptide ions by providing a second, 'mobile' proton, which lowers amide bond strength and allows more facile unimolecular decay (Dongré *et al*, 1996). N-terminal tags containing a sulfo group have been advocated as a useful approach to generate a full y-ion series of peptide fragments in MS/MS (Keough *et al*, 1999; Samyn *et al*, 2004; Wang *et al*, 2004). By simple manual calculation of the differences between the adjacent y-ion fragments, or by using suitable software, the amino acid sequence can readily be interpreted. As observed before, the introduction of a negative charge usually leads to a decrease or even loss in signal intensity in positive mode. However, when the MS analysis is performed in negative mode most of the signals are observed and the corresponding, not-observed, precursor (+2 Da) can be fragmented in positive MS/MS analysis.

Using the improved in-gel protocol, we were able to identify 53 2D-PAGE separated proteins from *Shewanella oneidensis*. As most of the steps in the protocol can be performed in-gel (guanidination, trypsin digestion and peptide extraction) this should allow a high-throughput approach on an automated platform. As demonstrated, the use of such platforms offers several advantages: it is less labor-intensive, simpler, faster and there is less potential exposure to lab contaminants (Houthaeve *et al*, 1997). Since the loss of intact proteins from a gel spot is minimal, this should also have a positive effect on the peak intensity during MS-

analysis. The proteins in the spots were identified by using the *de novo* determined sequences in a novel search algorithm. Searches using the *de novo* data were done against the NCBI-database, resulting in multiple hits for every search. In all spots, the hit with the best probability score was a protein from *Shewanella* whereas hits with lower scores corresponded to homologous proteins from related species. FASTS searches on sequence similarity and, therefore, this approach provides the possibility to identify proteins from organisms of which the genomes have not been sequenced.

3.3. In-gel guanidination; applications and automation.

In the previous chapter (Part 3.2) the development of the in-gel guanidination protocol was described along with a proof-of-principle study on 2D-PAGE separated proteins from the bacterium *Shewanella oneidensis*. However, the real benefit of this protocol is that the *de novo* determined sequence information can be used for homology-based identification of proteins from non-model organisms (Shevchenko *et al*, 2001). Therefore, we applied our protocol on proteins isolated from two distinct species having none or very limited nucleotide or amino acid sequence information. *Halorhodospira halophila* (Part 3.3.1) is an extreme halophilic phototrophic purple bacterium, studied in our laboratory because of its photophobic response towards blue light. The monocotyledon *Musa sp*. (banana), an important food crop in the humid tropics (Marriott, 1980), has been studied as second non-model organism (Part 3.3.2). The identification of proteins from this plant is situated in an collaborative effort, coordinated by Professor Swennen from the Laboratory of Tropical Crop Improvement (KUL), to study the possibilities of cryopreservation as a tool to preserve the biodiversity of banana-species.

The preparation of proteome samples is by far the most daunting exercise in current approaches. Here lies the main stumbling block for proteomics and consequently its high-throughput analytical capabilities. Today's high-throughput identification of gel-electrophoresed proteins heavily relies on automation. A single large-scale study based on 2D-PAGE can result in hundreds of spots to be identified by MS. To achieve this task in a reasonable time frame, parallel sample preparation and subsequently automated MS-analysis is essential (Quadroni *et al*, 1999). Furthermore, the use of such automated proteomic platforms offers several advantages: it is less labor-intensive, simpler, faster, and there is less potential exposure to lab contaminants (Houthaeve *et al*, 1997). The implementation of in-gel guanidination results in a simplification of the protocol for N-terminal sulfonation.

In the final chapter (Part 3.3.3), modifications to the in-gel guanidination protocol, to make it amendable to automation, and the automation of this protocol are described. The most important modification in the protocol is the use of an alternative sulfonation reagent. In the previous studies, we used 2-sulfobenzoic acid cyclic anhydride in THF. However, due to its extreme volatility, THF is difficult to use in automated approaches. In Part 3.3.3.1, the use of this reagent is compared to the use of the less reactive and water-compatible reagent 4-sulfophenyl isothiocyanate (SPITC) (Marekov *et al*, 2003). The automation, using an Ettan workstation (GE Healthcare), and initial results of the automated protocol are described in Part 3.3.3.2.

3.3.1. Application 1: *Halorhodospira halophila*

MALDI TOF/TOF *de novo* sequence analysis of 2D-PAGE separated proteins from *Halorhodospira halophila*, a bacterium with unsequenced genome

Bart Samyn, Kjell Sergeant, Samy Memmi, Griet Debyser, Bart Devreese and Jozef Van Beeumen

University Gent
Department Biochemistry, Physiology and Microbiology
Laboratory of Protein Biochemistry and Protein Engineering
K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

.

In press Electrophoresis (**2006**)

Introduction

Proteomics is playing a pivotal role in the post-genome era in helping to define the functional role of genes. Mass spectrometry (MS), hyphenated with a range of electrophoretic and multidimensional chromatographic separation techniques, has emerged as a key platform technology in proteomics for the rapid and high-throughput identification, characterization, and quantitation of proteins (Aebersold *et al*, 2003). Typically, proteins are digested using trypsin and the resultant peptides are then subjected to MS analysis. The tryptic peptides provide a characteristic mass fingerprint (PMF), which can be used to identify proteins. Although this approach is useful to identify proteins in simple mixtures, peptide sequence information obtained by tandem mass spectrometry (MS/MS) is required to identify individual proteins in more complex samples (Kapp *et al*, 2003). Sophisticated algorithms (e.g. SEQUEST, Mascot) have been developed to aid in this process, starting from peptide MS/MS data whereby peptides are identified by correlating the uninterpreted MS/MS spectra with simulated (predicted) product ion spectra derived from peptides of the same mass contained in the databases. While the above-mentioned algorithms for protein identification from peptide MS/MS data have enjoyed considerable success, their utility is directly related to the quality of the product ion spectra and depends on the availability of database information about the proteins under investigation. For proteins not contained within sequence databases, it is necessary to determine partial or complete amino acid sequences using either manual or automated *de novo* peptide sequence analysis methods.

Since manual *de novo* sequencing is a very time-consuming process, several software tools were developed that deduce an amino acid sequence from an MS/MS spectrum. Interpretation of MS/MS spectra relies on measuring the mass differences between adjacent fragment ion peaks of one of the major ion series, i.e. b-series (ions containing N-terminus) or y-series (ions containing C-terminus), which are common in tryptic peptides. However, most of *de novo* sequencing software tools inevitably suffers from inherent limitations of MS/MS spectral analysis, including incomplete b- and y-ion series (gaps), the presence of other peptide-derived peaks such as a-ions, internal fragments and neutral losses of water or ammonia (Grossmann *et al*, 2005). The recent introduction of MALDI TOF/TOF MS technology offers the advantage of MALDI ionization with tandem mass spectrometry in a time-of-flight instrument (Medzihradszky *et al*, 2000). *De novo* sequencing of underivatized peptides using MALDI TOF/TOF MS has recently been demonstrated (Yergey *et al*, 2002). However, compared to the MS/MS spectra of doubly charged ESI-generated ions, MS/MS spectra of singly charged MALDI ions contain more ions from other fragment ion series. Therefore, a number of derivatization methods have been proposed to improve the fragmentation of singly charged ions.

One approach to facilitate the interpretation of MS/MS spectra is to enrich a series of fragments by attaching a strongly negatively or positively charged group to the N-terminus of peptides. Keough *et al* demonstrated that the N-terminus can be derivatized by acylation with 2-sulfobenzoic acid cyclic anhydride or chlorosulfonylacetyl chloride (Keough *et al*, 1999). N-terminal sulfonic acid derivatives were subsequently proposed for peptide sequencing by ESI MS, MALDI-TOF MS (Keough *et al*, 2003) and MALDI TOF/TOF MS (Samyn *et al*, 2004). The introduction of a sulfo group facilitates the MS/MS fragmentation of singly charged peptide ions by providing a second 'mobile' proton which lowers amide bond strength and allows more facile unimolecular decay (Dongré *et al*, 1996). Since sulfonation reagents react with amino groups, this derivatization results in the modification of both the N-termini and the ε-amines of lysine-containing peptides. Therefore, Keough *et al*. expanded

their approach and combined guanidination of lysine residues with the addition of a sulfonic acid group to the N-terminus (Keough *et al*, 2000b). Following guanidination of lysine ε-amines, introduction of sulfonic acid groups to tryptic peptides is possible solely at the N-terminus. We further improved this method by performing the Lys side-chains modification directly on gel-separated proteins prior to tryptic digestion. In this way, removal of the molar excess of guanidination reagents can simply be accomplished during the destaining step of the gel spots (Sergeant *et al*, 2005).

Today, most proteomic studies of extremophilic bacteria have been performed on members of the Archaea for which a sequenced genome is available. Zhu *et al* applied the MudPIT approach to analyze the proteome of *Methanococcus jannaschii*, an autotrophic methanoarchaeon and the first member of the Archaea with a completely sequenced genome (Zhu *et al*, 2004). A shotgun approach was also used to identify the *Sulfolobus solfataricus* P2 proteome, a thermo-acidophilic crenarcheon (Chong *et al*, 2005). The cytosolic and membrane proteome of *Halobacterium salinarum* has been analyzed using PMF and LC-MS/MS techniques (Klein *et al*, 2005; Tebbe *et al*, 2005). *Halobacterium salinarum* is a member of the halophilic archaea and an important model organism to study adaptations necessary for living in salty habitats. The genome sequence of *Halobacterium* species NRC-1 has completely been determined (Ng *et al*, 2000).

The extremely halophilic purple phototrophic bacterium *Halorhodospira* (formerly *Ectothiorhodospira*) *halophila* shows a photophobic response towards intense blue light. The wavelength dependence of this response corresponds with the absorption spectrum of the photoactive yellow protein (PYP), which suggests this protein to be the primary photoreceptor for this response (Sprenger *et al*, 1993). Photoreceptors allow living organisms to make optimal use of the light conditions for growth and development and/or the protection from light damage. Various types of light-induced sensory responses have been characterized physiologically in detail. However, the molecular basis of this type of response is only slowly emerging. While the PYP protein is extremely well studied at the physical level, direct proof of a link between PYP and negative phototaxis is lacking. Moreover, in other species that produce PYP, a link with phototaxis has never been reported. A major limitation to study the physiological function of PYP is the fact that sequence information about the genome is not available (a DOE funded sequence program is currently running) and, more in particular, genetic techniques are poorly developed in this organism. There is limited information about the flanking regions of the PYP gene (Kyndt *et al*, 2005) in *H. halophila* and other species, but except for the presence of the biosynthetic genes for the production of the co-factor (p-coumaric acid) and for its covalent attachment to the protein, there are no generalizations that provide clues concerning the role of PYP.

Here, as a proof of principle, we applied our improved MS identification approach to identify a number of 2D-PAGE separated proteins from *Halorhodospira halophila*. (Partial) sequences of tryptic peptides were submitted to homology-search for identification of the corresponding protein. For this purpose, we applied three different homology-based search algorithms, MS-Homology, FASTS and MS BLAST (Clauser *et al*, 1999; Shevchenko *et al*, 2001; Mackey *et al*, 2002).

Materials and methods

Materials

Urea, ammonium persulfate, CBB G-250 and agarose were obtained from Amersham Biosciences (Uppsala, Sweden). Iodoacetamide, CHAPS, DTT and TEMED were from Fluka (Buchs, Switzerland). Immobilized pH gradient (IPG) strips, SDS, glycine and ampholytes were purchased from Bio-Rad (Hercules, CA, USA). The acrylamide/bisacrylamide solution was obtained from National Diagnostics (Atlanta, Ge, USA), and the solvents for mass spectrometric sample preparation were from Biosolve (Valkenswaard, The Netherlands).

Bacterial growth and preparation of extracts

Halorhodospira halophila SL-1 was grown anaerobically under tungsten illumination at 30° in medium 253 described by DSMZ, and 2.5 mL of these cultures was used to inoculate 250 mL anaerobically prepared medium. Cultures were grown anaerobically under tungsten illumination or green and blue light conditions at 30°C and harvested at the late-exponential growth phase. After washing with dH_2O, the cells were resuspended in 100 mM Tris-HCl, pH 8.0, supplemented with 50 µg DNAse and 0.5 mM PMSF, and fractionated by sonication, followed by centrifugation to remove the cell debris.

Two-dimensional gel electrophoresis and analysis

After determination of the protein concentration with the Protein Assay Kit (Bio-Rad), approximately 250 µg of protein was mixed with IPG rehydration buffer (8M ureum, 2% w/v CHAPS, 0.3% DTT, final volume = 360 µL). The strips were allowed to rehydrate for 7h and to focus (IEF) using a Multiphor II system (Amersham Biosciences) running the following program: 150 V (30'), 150 V (120'), 300 V (30'), 300 V (45'), 3500 V (90'), 3500 V (540'), 500 V (10') and hold at 500V. The temperature was kept at 18°C. After completion of the IEF program, the IPG strips were equilibrated in a 50 mM Tris-HCl solution, pH 8.8, containing 6 M urea, 30 % glycerol, 2% SDS and 1% DTT, for 10 min, after which the solution was replaced with the same solution, except that DTT was exchanged by 5% iodoacetamide. The strips were then placed on the home-casted vertical SDS-PAGE gels and subjected to electrophoresis at 10 mA/ gel for 15 min, followed by a +/- 5 h run at 20 mA/gel until the Bromophenol Blue front reached the bottom of the gel. Staining was performed using CBB G-250. The 2D-gel images were digitized using a GS-710 densitometer (Bio-Rad) and analysed with the accompanying PDQuest 7.1 software (Bio-Rad).

In-gel guanidination

Guanidination was performed by adding 5 µl MQ, 11 µl 7 N ammonium hydroxide (Merck, Darmstadt, Germany) and 3 µl of a freshly prepared 7.5 M *O*-methylisourea hemisulfate solution (Across, Geel, Belgium) to the excised spots. The samples were vortexed briefly and incubated at 65°C. After incubation for two hours, the guanidinated samples were taken from the oven and the remainder of the solution was discarded. The gel pieces containing the guanidinated samples were desalted and destained in one step. Two washes using 150 µl 200 mM ammonium bicarbonate in 50 % ACN/MQ (30 minutes at 30°C) were performed and subsequently the gel pieces were dried in a SpeedVac (Thermo Savant, Holbrook, NY).

Trypsin digestion and sulfonation

A volume of 8 µl digestion buffer (50 mM ammonium bicarbonate, pH 7.8) containing 150 ng modified trypsin per µl (Promega, Madison, WI) was added to the dried gel spots and the tubes were kept on ice for 45 minutes to allow the gel pieces to be completely soaked with the protease solution. Digestion was performed overnight at 37°C, the supernatans was recovered and the resulting peptides were extracted twice with 35 µl 60% ACN/0.1% DIEA. The extracts were pooled and dried in the SpeedVac. The peptides were redissolved in 4 µl 12.5 mM ammonium bicarbonate 50% ACN/MQ, and 2 µl was mixed with 2 µl of the sulfonation solution. The sulfonation reagent was prepared by dissolving 2 mg 2-sulfobenzoic acid cyclic anhydride in 1 ml dry THF to attain a 0.01 mM solution. The tubes were briefly vortexed and reacted for 15 minutes at room temperature.

MS and MS/MS

A 4700 Proteomics Analyzer (Applied Biosystems, Foster City, CA) with TOF/TOF optics was used for all MALDI MS and MS/MS applications. Samples were prepared by mixing 0.7 µl of the sample with 0.7 µl matrix solution (7 mg/ml CHCA in 50% ACN containing 0.1% TFA) and spotted on a stainless steel 192-well target plate. They were allowed to air-dry at room temperature, and were then inserted in the mass spectrometer and subjected to mass analysis. The mass spectrometer was externally calibrated with a mixture of Angiotensin I, Glu-fibrino-peptide B, ACTH (1-17), and ACTH (18-39). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrino-peptide B.

All of the sulfonated peptides were subjected to MS/MS using a MALDI TOF/TOF instrument. In an initial study, using this method, in which the fragmentation spectra resulting from high- and low-energy CID experiments were compared, the authors concluded that the difference in fragmentation and the effect on database search results was surprisingly small (Sumpton *et al*, 2003). The major difference observed is the presence of high-energy fragment ions (w-ions) in the high-energy CID spectra of some peptides. When the collision induced dissociation mode (gas on, collision energy 0.5 to >3.5 keV) is applied, a larger number of low molecular weight fragments (immonium ions, internal fragments) have been observed (Walker *et al*, 2003). However, it has also been shown that the use of high energy CID results in a loss of sequence information, as the y-ion abundance decreases at both higher gas pressure and higher collision energy (Campbell, 2003). Therefore, we performed all fragmentation experiments with the collision energy set at 1 keV and no gas in the collision chamber (low-energy CID).

Database searches

The *de novo* determined peptide sequences were deduced manually and used for similarity searches using the FASTS, MS BLAST and the MS-Homology algorithm. On-line submissions were performed using MS BLAST at the Heidelberg server (http://dove.embl-heidelberg.de/Blast2/msblast.html). Searches were performed against the non-redundant database (nrdb) using standard settings. The FASTS algorithm (http://fasta.bioch.virginia.edu/fasta_www/cgi/) was applied using standard settings, and searches were performed against the NCBI/BLAST nrdb with BLOSUM 50 as search matrix. MS-Homology searches (Protein Prospector 4.0.5) were performed on the UCSF server

against the NCBI nrdb using BLOSUM 50 as search matrix (http://prospector.ucsf.edu /ucsfhtml4.0/mshomology.htm).

The software used for similarity searches does not discriminate between the isobaric amino acids Ile and Leu. Therefore, all mass increments of 113 Da between consecutive y-ions were arbitrarily designated as Ile. The FASTS search results were considered significant if the E-value was below 1.0 e-04. The MS BLAST search results were considered significant if the resulting scores were higher than the threshold score indicated in the software. In order for a particular protein in the database to generate a hit, MS-Homology must find homologous sequences for the minimum number of peptides required to match. The scoring method used is based on a mutation matrix such as the one used in the BLAST and FASTA programs. The final score is calculated by adding the scores for the individual peptide alignments together. If there are several possible alignments of a given peptide, then the highest scoring alignment is used in the calculation. As the searches are based on similarity, proteins identified with lower scores must have the same generic function as the first hit. Proteins were considered positively identified only if all three search algorithms yielded the same homologous protein in the first hit. It has been demonstrated that indirect evidence can add to the significance of an identification (Shevchenko *et al*, 2001). Therefore, the identifications were further validated by using information such as the cleavage specificity of trypsin and sequence information resulting from known preferential fragmentation patterns of sulfonated peptides (Samyn *et al*, 2004).

Results and discussion

The total protein extracts from *Halorhodospira halophila* grown anaerobically under yellow or green/blue light were separated by 2D-PAGE. From these two gels, 100 spots were randomly selected and manually excized. The proteins were guanidinated in-gel and desalted/destained in one single step as described previously. Subsequently, the guanidinated proteins were enzymatically cleaved with trypsin and, after extraction, the peptides were sulfonated (Sergeant *et al*, 2005). For 74 spots, we observed a good PMF of the sulfonated peptides, suitable for *de novo* MALDI MS/MS analysis. For the other 26 spots, we observed none or a very weak PMF, with signal intensities that were too low for MALDI MS/MS fragmentation. Previous experiments have indicated that sulfonic acid-derivatized peptides have poorer positive-ion sensitivity than the corresponding native peptides. The introduction of a negative charge usually leads to a decrease or even loss in signal intensity in positive mode (Samyn *et al*, 2004). Keough *et al* also noticed a decreased intensity of sulfonated peptides in the positive mode, compared to the negative ion mode and, apparently, some peptides show no signal above the noise level (Keough *et al*, 1999). Recently, it was also demonstrated that the presence of the strong negative charge of the sulfonic group can create problems for sample desalting on reversed-phase media (low yield for less hydrophobic peptides) (Raucci *et al*, 2005). However, in our approach, guanidination is performed in-gel, and therefore, an additional desalting step to remove the excess of reagents, can be omitted (Sergeant *et al*, 2005).

All fragmentation experiments were performed with the collision energy set at 1 keV and no gas in the collision chamber (low-energy CID). Typically, the most intense peaks in the PMF were selected for MS/MS analysis. In most spots, three to six peptide sequences, with a length varying between five and twenty amino acids, were obtained *de novo* (Table 3.7a & b). In all fragment spectra an initial loss of the sulfonic acid derivative was observed (Δm = 184 Da). By simple manual calculation of the differences between the adjacent y-ion

fragments, the amino acid sequence could readily be interpreted. As an example, the protein in spot 5 was identified as fructose-1,6-bisphosphate aldolase (Table 3.7a). Upon sulfonation, four peptides were subjected to MALDI MS/MS analysis. In all fragmentation spectra, except one, we observed a complete y-ion series that could easily be interpreted, facilitating *de novo* sequencing (Figure 3.20a-d). MS/MS spectra of derivatized peptides having an internal homo-Arg (guanidinated lysine) also produced complete y-ion series (Figure 3.21a). However, the y-ion series were often accompanied with (y_i –17)-ion series, due to the neutral loss of NH_3. Although this (y_i –17) series is seen in the fragment spectra up to the step where the basic residue is cleaved off, its presence does not interfere with the sequence interpretation (Figure 3.20a-d). Please note that the loss of the sulfonic group (-184 Da) and the neutral loss of ammonia (-17 Da) is summed as an initial loss of –201 Da.

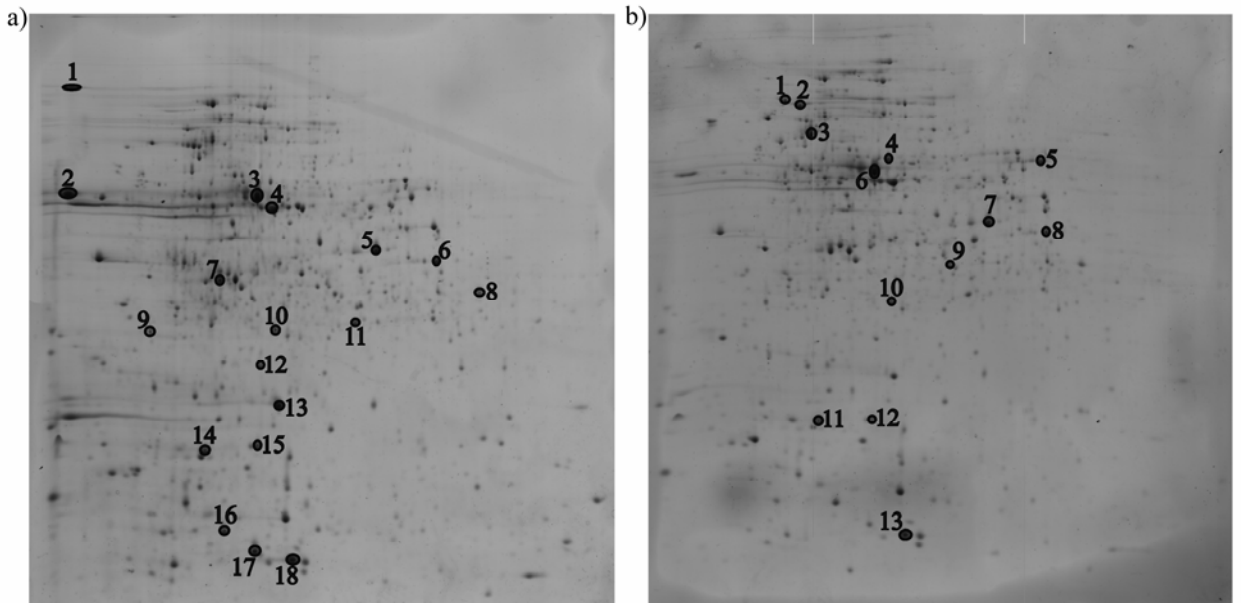


Figure 3.19. 2D-PAGE separated proteins from *Halorhodospira halophila* grown anaerobically under tungsten illumination (a) or green and blue light (b). Spots in which the protein was positively identified are numbered according to Table 3.7.

Table 3.7a. Identified proteins from *Halorhodopira halophila (*yellow light)

spot[a]	protein[b]	Identification FASTS[c]	Pept.[d]	AA[e]	E-score[f]	MS BLAST score[g]	MS-Hom. score[h]
1,2	71909086	porin, Gram-negative type	5	50	3.50e-05	135	124
3	37527547	S-adenosylmethionine synthetase	6	54	2.20e-17	202	216
4	39936346	elongation factor Tu	6	73	2.90e-29	336	376
5	26991638	fructose-1,6-biphosphate aldolase	4	32	6.60e-14	115	186
6	33634721	phosphoribulokinase	4	33	5.40e-05	95	129
7	9392587	sarcosine-dimethylglycine methyltransf	4	31	0.00021	-	113
8	34497819	acetoacetyl-CoA reductase	4	51	4.30e-10	149	159
9	74317158	triosephosphate isomerase	3	29	7.80e-12	156	121
10	34498798	adenylate kinase	3	31	1.10e-15	177	191
11	53804425	2.,3,4,5-tetrahydropyridine-2,6-di-carboxylate N-succinyltransferase	5	45	1.40e-08	167	144
12	47574096	pentose-5-phosphate-3- epimerase	3	29	0.00046	105	98
13	67941974	superoxide dismutase	8	67	5.70e-17	313	298
14	68304953	DsrC	1	13	1.60e-07	99	82
15	78700374	nucleoside diphosphate kinase	3	49	4.00e-29	279	259
16	132132	ribulose biphosphate carboxyl small chain	3	34	1.10e-06	107	113
17	53805138	pterin-4-α-carbinolamine dehydratase	4	64	4.00e-13	188	132
18	69951812	cold-shock protein, DNA-binding	2	21	2.70e-08	112	115

Table 3.7b. Identified proteins from *Halorhodospira halophila (*green/blue light)

spot[a]	protein[b]	Identification FASTS[c]	pept.[d]	AA[e]	E-score[f]	MS BLAST score[g]	MS-Hom. score[h]
1	77166263	chaperone protein dnaK (Hsp70)	4	35	1.60e-11	186	155
2	54294307	30S ribosomal protein S1	3	34	1.60e-13	152	181
3	53762519	chaperonin GroEL (HSP60 family)	6	83	8.60e-26	383	299
4	71899446	ATP synthase F1, alpha subunit	2	35	1.00e-10	152	127
5	71550918	ribulose-bisphosphate carboxylase	4	30	1.60e-05	99	160
6	37527547	S-adenosylmethionine synthetase	6	56	1.00e-22	194	257
7	56461311	fructose/tagatose bisphosphate aldolase	5	29	4.40e-11	109	156
8	33862805	phosphoribulokinase	3	36	1.40e-16	158	191
9	33152351	NADH- dependent enoyl-ACP reductase	4	40	1.20e-12	161	190
10	2497482	adenylate kinase (ATP-AMP transphosphorylase)	4	37	3.1e-09	174	195
11	52006362	dissimilatory sulfite reductase	1	13	0.00097	73	70
12	77866837	nucleoside diphosphate kinase	2	24	4.70e-10	101	132
13	1402737	major cold shock protein	1	11	9.70e-05	77	71

[a] Spot number according to the position on the 2D-PAGE (Figure 3.19)
[b] NCBI *Entrez* entries (http://www.ncbi.nih.gov/Entrez/)
[c] Identification based on FASTS search result
[d] Number of peptide sequences used in the query
[e] Total number of amino acids used in the query
[f] In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI non-redundant protein database N ≈ 2075116.
[g] MS BLAST score for searches against the non-redundant protein database at http://dove.embl-heidelberg.de/Bla st2/msblast.html.
[h] MS-Homology (Protein Prospector 4.0.5) score against the NCBI non-redundant protein database.

During post source decay (PSD) experiments with derivatized peptides enhanced fragmentation at Pro residues and a reduced abundance in fragmentation at the C-terminal side of Pro has been observed (Klein *et al*, 2005). Here, and in a previous study, TOF/TOF analysis of peptides containing an internal Pro indicated that y-ions resulting from cleavage on the N-terminal side of Pro are enhanced while y-ions resulting from cleavage on the C-terminal side of Pro are less abundant or almost completely depleted (Samyn *et al*, 2004). Therefore, peptides with a Pro residue near the N-terminus of a peptide can limit the amount of sequence information and cause gaps in the derived sequence (Figures 3.20d & 3.21c). If the Pro residue occurs in the middle or near the C-terminus of the peptide this effect is less detrimental (Figure 3.20c), and even for derivatized peptides containing internal homo-Arg and Pro residues we were able to derive uninterrupted peptide sequences of 20 amino acids (Figure 3.22b & c). As proteomic strategies are becoming increasingly reliant on the use of automated database search algorithms, incorporation of 'fragmentation rules', such as the observed preferential cleavage at Xxx-Pro peptide bonds, into the database search algorithms will aid in the development of more effective tools for high-throughput protein identification. Furthermore, the occurrence of 'non-sequence' specific ion fragments, such as the neutral loss of ammonia from peptides with internal Arg or homo-Arg, can be used to improve predictive models of peptide fragmentation for *de novo* sequence analysis.

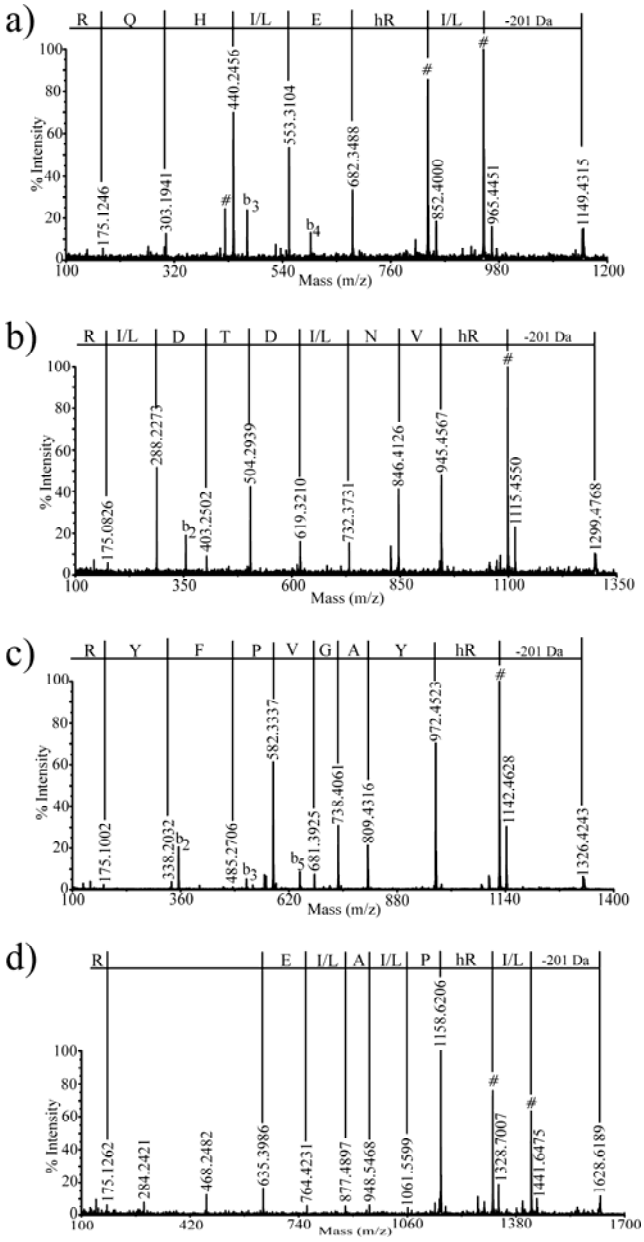


Figure 3.20. MALDI MS/MS spectra (positive ion mode) of the in-gel guanidinated and sulfonated tryptic peptide mixture of spot 5 (Figure 3.19). Fragment spectra of peptide IKEIHQR (a), KVNIDTDIR (b), KYAGVPFYR (c) and IKPIQIE (d). All labeled fragment ions are y-ions, (y-17)-ions resulting from the neutral loss of NH_3 are indicated as #. Where appropriate, other fragment ions are indicated (b-ions). The de novo derived sequence information is indicated in the one-letter code (hR, homoarginine). The loss of the sulfonation label is indicated as -184 Da or as a loss of -201 Da including the neutral loss of ammonia (-17 Da).

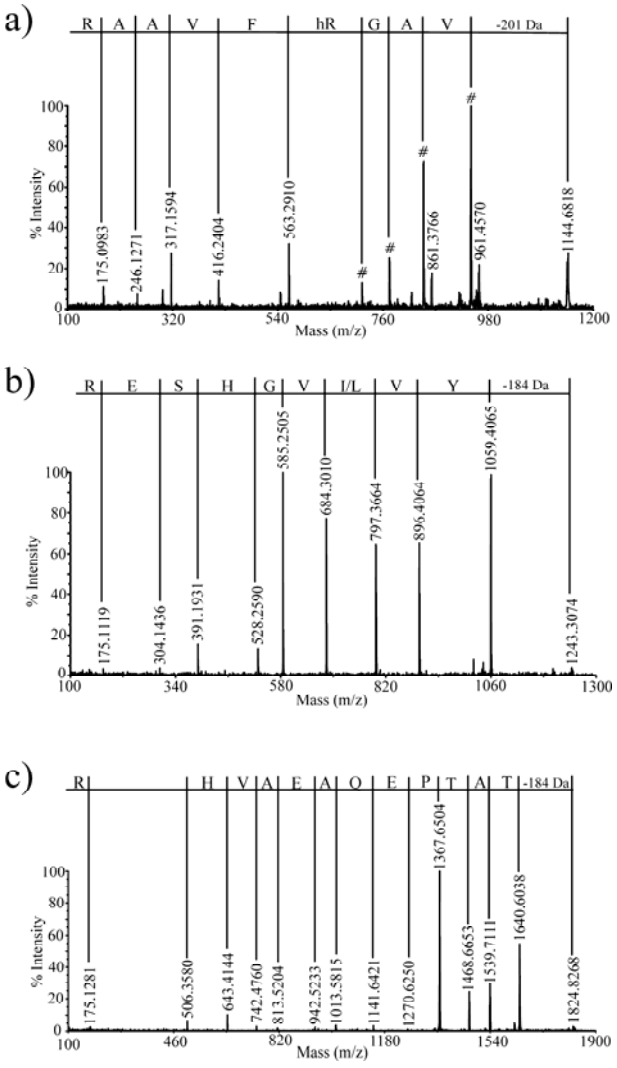


Figure 3.21. MALDI MS/MS spectra of the in-gel guanidinated and sulfonated tryptic peptide mixture of spot 9 (Figure 3.19). Fragment spectra of peptide VAGKFVAAR (a), YVIVGHSER (b) and TATPEQAEAVH (c). Labeling is as in Legend Figure 2.

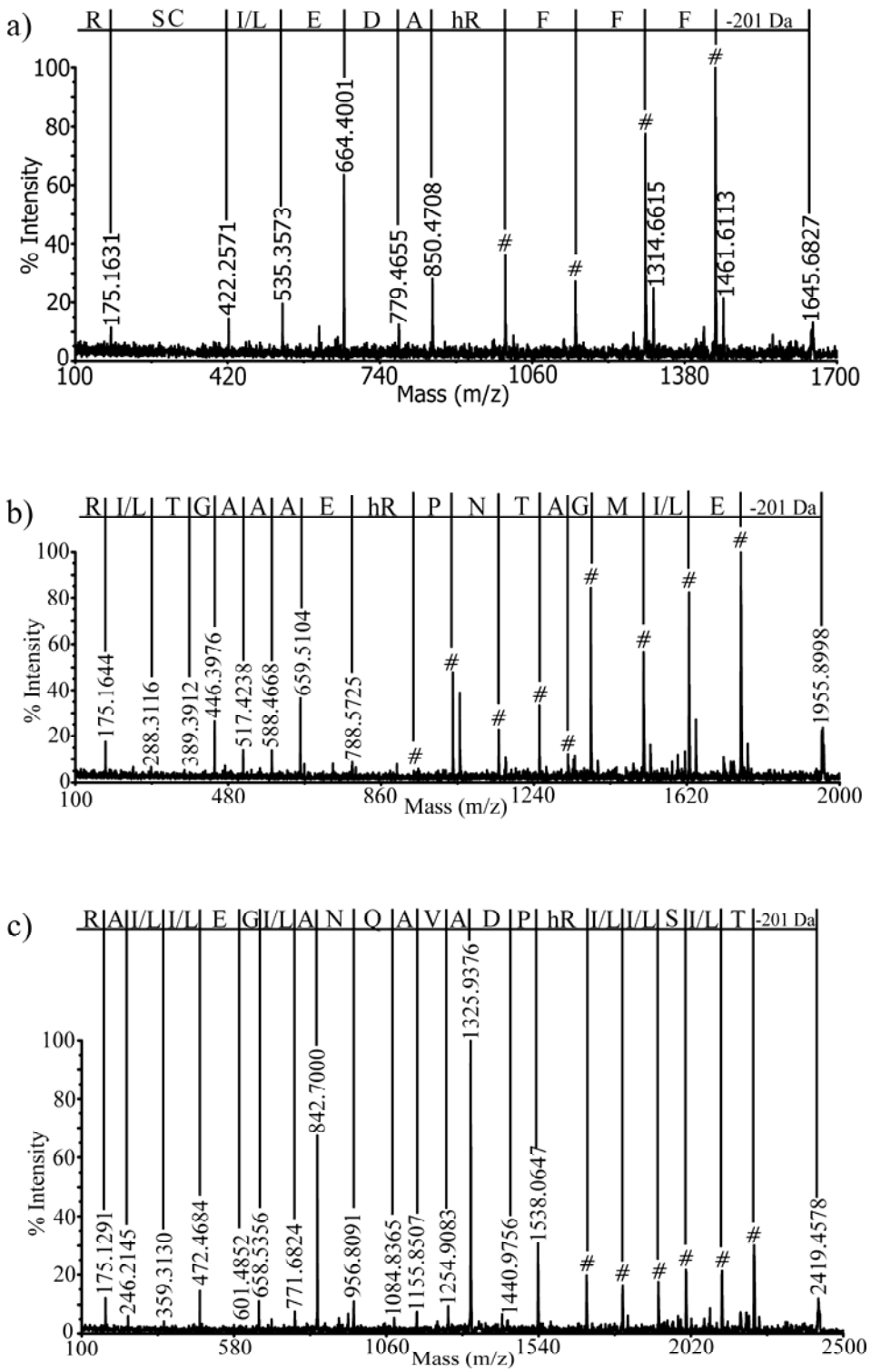


Figure 3.22. MALDI MS/MS spectra (positive ion mode) of the in-gel guanidinated and sulfonated tryptic peptide mixture of spot 15 (Figure 3.19). Fragment spectra of peptide FFFKADEIFTR (a), EIMGATNPKEAAAGTIR (b) and TISIIKPDAVAQNAIGEIIAR (c). Labeling is as in Legend Figure 2.

The *de novo* determined peptide sequences were used to identify the proteins by sequence similarity searching, as was recently reviewed (Liska *et al*, 2003a). Given the size and growth of the current databases, it is possible that many proteins already have homologues in a database. Database searching with MS-derived *de novo* peptide sequences allows the proteomic identification of proteins from organisms whose genomes have not been sequenced. However, MS and sequence similarity searches are difficult to combine.

Conventional database search algorithms like BLAST or FASTA are optimized for accurate sequence queries that are longer than 35 amino acid residues. Usually peptide sequences obtained by MS/MS do not exceed the length of a tryptic peptide, typically comprising 10-15 amino acids and, therefore, the statistical significance of retrieved hits is often ambiguous. Several database searching approaches have been reported that accommodate specific requirements of MS/MS sequencing. Mass spectrometry-driven BLAST (MS BLAST) is a database search protocol for identifying unknown proteins by sequence similarity to homologous proteins available in a database. MS BLAST utilizes redundant, degenerate, and partially inaccurate peptide sequence data obtained by *de novo* interpretation of MS/MS spectra. MS BLAST does not allow gaps within individual peptides, while gaps between peptides are not penalized and can be of arbitrary length. Therefore, all peptide sequences obtained by the interpretation of acquired MS/MS are assembled into a single searching string in arbitrary order (Shevchenko *et al*, 2001; Habermann *et al*, 2004). MS-Homology is a database searching tool from the UCSF Mass Spectrometry Facility (Protein Prospector 4.0.5) that performs homology-based searches (Clauser *et al*, 1999). The program allows to compare a number of *de novo* derived peptide sequences, followed by the maximum number of amino acid substitutions allowed for each sequence, against a selected database. Different peptides from the same unknown protein can be entered in the list. A database search will look for proteins containing peptides identical or homologous to the listed sequences. The quality of the results will be dependent on the number of peptides sequenced and the accuracy of the sequence information entered, as well as on database completeness and species to species sequence variability for the peptides entered. It is also possible to enter a part of the sequence as a mass, along with a tolerance factor. FASTS is a recently reported sequence similarity search algorithm designed to use *de novo* sequence data from organisms lacking comprehensive sequence data. FASTS searches databases using peptide sequences of unknown order, evaluating all possible arrangements of the peptides. The algorithm uses the heuristic FASTA comparison strategy to accelerate the search, but also uses alignment probability, rather than a similarity score, as the criterion for alignment optimality. Because the true order of the query peptides used by FASTS is not known, FASTS only requires that the aligned peptides do not overlap (Mackey *et al*, 2002).

The *de novo* derived sequence information from each spot was combined in one search query and analyzed using the three search algorithms. Only when the top results (first hits) from the three searches yielded the same protein, the identification was considered as positive. Most search queries included 30 to 70 amino acids, resulting from three to six peptide sequences (Appendix IIa & b). Using these queries, all three homology-based search algorithms yielded identifications with a score significantly better than the threshold score (Table 3.7a & b). Three proteins were identified using only two *de novo* derived peptide sequences, and another three proteins were identified by using only one peptide sequence (> 10 amino acid residues). In the latter cases, although all three algorithms yielded the same homologous protein, confirming a positive identification, the identification scores dropped significantly (Table 3.7a & b). Surprisingly, we observed that the identification score obtained varied slightly according to the position of the peptide sequences in the query using MS BLAST and, to a lesser extent, using the FASTS search algorithm. As it was formerly suggested that the peptide sequences can be used in an arbitrary way, this observation will be the subject of further investigation.

Using this approach, we were able to identify 31 proteins in the 74 spots, in which a PMF was observed, from both gels (42 %) (Table 3.7a & b). In some of the fragmentation spectra none or insufficient sequence information was derived, most likely because of the

weak intensity of the derivatized precursor. For other spots, sufficient *de novo* sequence information was obtained, but the homology search algorithms yielded a protein identification with a score below the threshold value, or no identification at all. According to simulation results, sequence-based methods, such as MS BLAST and FASTS, are able to detect +/- 50% of homologous sequences at the sequence identity level of +/- 50%. The success of identification by sequence similarity searches will also depend on the number of recognized peptides from a digested protein. It has been calculated that, as more peptides are analyzed and matched, proteins of less similarity can be identified, the limit being around 50% identity. The simplest cases are those in which the proteins in question are highly conserved and can thus be identified via the sequences of their homologous proteins in other species. This strategy fails when the proteins are insufficiently similar. If the organism being studied is very distantly related to any organism with a sequenced genome, the likelihood of protein identification decreases (Mackey *et al*, 2002; Sunyaev *et al*, 2003).

Some of the proteins we identified are definitively involved in the adaptation to halophilic life conditions. Sarcosine dimethylglycine N-methyltransferase (SDMT) of *Ectothiorhodospira halochloris* catalyzes the threefold methylation of glycine to betaine, with S-adenosylmethionine (SAM) acting as the methyl group donor. Glycine betaine is accumulated in cells living in high salt concentrations in order to balance the osmotic pressure. SAM has an important role in DNA methylation and cell signaling. S-adenosylmethionine synthetase catalyzes the formation of S-adenosylmethionine from L-methionine and ATP. However, a complete study of the photorespons of *H. halophila*, and the role of PYP in this process, will require a differential display study of 2D-PAGE separated proteins from cells grown under different light conditions.

Concluding remarks

The rapid and accurate identification of proteins is the primary goal of modern proteomics. Tandem mass spectrometry (MS/MS) can generate some useful sequence information. However, manual interpretation of peptide spectra for de novo sequencing is often prohibitively challenging because of variation in favored ion fragmentation sites, the chemical nature of amino acid side chains and their relative order in a peptide backbone, and the presence of side-products such as neutral loss ions, contaminants, or noise peaks. Improvement of the fragmentation efficiency of peptides is of particular importance for MALDI-generated ions, because the predominant singly charged ions in MALDI generally fragment less good than doubly charged ions. The approach demonstrated here, consisting of de novo sequence analysis of derivatized peptides and homology-based identification, is a powerful technique for the identification of proteins with no genomic or other database information.

For 75 % of the spots we observed a PMF upon in-gel guanidination and sulfonation of the extracted peptides. An apparently bad feature of the sulfonic acid derivatized peptides is their lower intensity in positive mode analysis, partly due to the suppression effect of the strong negative charge from the sulfonic group. The poorer positive ion-sensitivity is counterbalanced by a far more efficient fragmentation of sulfonated compared with non-sulfonated peptides. TOF/TOF analysis of underivatized peptides typically results in complex fragment spectra. After guanidination and sulfonation, a contiguous series of y-ions was observed in almost all of the fragmentation spectra (Appendix IIa & b). The y-ion series could easily be interpreted (manually or by using an algorithm) facilitating *de novo* sequencing.

The occurrence of 'non-sequence' specific ion fragments, such as the neutral loss of ammonia, and preferential fragmentation pathways, such as the Xxx-Pro bond, can be used to improve predictive models of peptide fragmentation for *de novo* sequence analysis. The current understanding of the fragmentation mechanisms is still insufficient to ensure a high correlation between theoretically predicted MS/MS spectra and experimental results.

MS-Homology, MS BLAST and FASTS methods provide independent means of evaluating the statistical significance of hits and, therefore, it is not necessary to compare retrospectively the matched peptide sequences with actual tandem mass spectra to rule out false positive hits. As previous reported, the peptide sequences in the query were arbitrarily chosen (Mackey *et al*, 2002; Sunyaev *et al*, 2003). However, in this study we observed that the identification scores vary, according to the position of the sequences in the query, when applying the MS BLAST and FASTS algorithm, a contradiction that will be the subject of a further study.

Note afterwards

The publication on 07/01/2006 of a draft of the *Halorhodospira halophila* genome (www.jgi.doe.gov/sequencing/DOEmicrobes2005.html), a few days after the article describing our work on this organism was submitted, allowed us to assess the performance of our protocol. Although not all proteins are represented in this draft sequence, those proteins that are represented allowed us to estimate whether the sequences we determined are correct. For instance the determined peptide sequence GVIKVGEEVEIVGITDTR, spot 4 on the gel of *Halorhodospira* grown under yellow light (Appendix IIa), is identical to the translated sequence. A BLAST search with this sequence against the entire NCBI-database revealed that this exact sequence does not occur in any known protein sequence. The closest match was found in translation elongation factor Tu from *Desulfovibrio vulgaris* (gi: 46581324) with the sequence GVIKVGEEVEIVGI**K**DT**TK**. Similar results were obtained for other peptides from this and other spots.

3.3.2. Application 2: *Musa* spp

Homology-based functional proteome analysis: a successful approach for the non-model plant *Musa* spp.

Bart Samyn[1]#, Kjell Sergeant[1]#, Sebastien Carpentier[2], Griet Debyser[1], Bart Panis[2], Rony Swennen[2] and Jozef Van Beeumen[1]

[1] University Gent
Department Biochemistry, Physiology and Microbiology
Laboratory of Protein Biochemistry and Protein Engineering
K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

[2] K.U. Leuven
Department of Biosystems
Laboratory of Tropical Crop Improvement
Kasteel Arenberg 13, 3001 Leuven, Belgium

# These authors contributed equally

Submitted for review J Prot Res

Introduction

The rapid and accurate identification of proteins is a primary goal of modern proteomics. Various mass spectrometric techniques, including MALDI-TOF and LC-ESI, can quickly obtain peptide mass maps that may be matched against theoretical spectra derived from primary sequence databases. Masses of intact tryptic peptides alone (in peptide mass fingerprinting, PMF), or together with masses of derived fragment ions using tandem mass spectrometry (MS/MS), are correlated with corresponding masses calculated by *in silico* processing of sequences from database entries. Sophisticated algorithms, such as Mascot (Perkins *et al*, 1999) and SEQUEST (Eng *et al*, 1994), have been developed to identify proteins from peptide MS/MS data. Peptides are identified by correlating the uninterpreted MS/MS spectra with simulated (predicted) product ion spectra derived from peptides of the same mass present in the available databases. While the above-mentioned algorithms for protein identification from peptide MS/MS data have enjoyed considerable success, their utility is directly related to the quality of the product ion spectra. Most commonly, both N- and C-terminal fragment ions are formed by most activation methods, and distinguishing them is not straightforward. Thus, if product ion spectra are formed that are not readily interpretable, low or insignificant search scores can result. The algorithms require an exact matching of analyzed peptides to database sequences and therefore, any discrepancy between sequences of the analyzed peptides, and between sequences of the corresponding database entries, typically results in mismatches and precludes the identification of the protein. Finally, for proteins not contained within sequence databases (it be a genome, an EST, or a protein sequence database), it is necessary to determine partial or complete amino acid sequences using either manual or automated *de novo* peptide sequence analysis methods.

Methods are needed that can extract protein-identifying information directly from spectra without comparison to databases. To achieve this goal a number of different de novo sequencing approaches have been developed in recent years. Several groups have taken advantage of the isotopic labeling approach. In addition to serving as mass-coded tags for relative quantification, such approaches also provide mass signatures for N- and C-terminal fragment ions and facilitate de novo sequencing. For example, proteolytic $^{18}O/^{16}O$ labeling has been used to code peptide C-termini in the course of tryptic digestion. As a result, y-type fragment ions are formed as mass-separated pairs that could easily be distinguished from b-ions (Shevchenko *et al*, 1997; Qin *et al*, 1998). On the other hand, attempts to simplify the fragmentation pattern have largely concentrated on the introduction of charged groups at the N- or C-terminus of the peptide. Keough and co-workers, e.g., developed a strategy in which peptide N-termini are derivatized by acylation in anhydrous medium with chlorosulfonylacetyl chloride (Keough *et al*, 1999). The addition of sulfonic acid groups to the N-termini of tryptic peptides facilitates de novo peptide sequencing by post source decay (PSD) MALDI-TOF MS. These derivatives promote efficient charge-site initiated cleavage of backbone amide bonds and allow the selective detection of only a single series of fragment ions that contain the original C-terminus of the peptide (y-ions).

MALDI TOF analysis, in the ion-reflection mode, is the most commonly used technique in proteomics because of its low-femtomol sensitivity and high resolution. MALDI sources predominantly ionize peptides as singly charged ions. To fragment these ions, higher collision energy is required and, therefore, cleavage of the amide bonds in the peptide backbone occurs less consistently. MALDI MS/MS spectra do not contain continuous ion series that facilitate the confident determination of long peptide sequences. With the availability of commercial instrumentation for MALDI tandem mass spectrometry, the use of

MALDI MS/MS in peptide sequence analysis has become a valuable tool for proteome analysis (Yergey *et al*, 2002). N-terminal sulfonic acid derivatives were proposed for peptide sequencing by MALDI-TOF MS (Keough *et al*, 1999; Keough *et al*, 2003) and MALDI TOF/TOF MS (Samyn *et al*, 2004). The sulfo group facilitates MS/MS fragmentation of singly charged peptide ions by providing an additional "mobile" proton, the acidic proton from the sulfo group, which lowers amide bond strength, allowing facile unimolecular decay (Dongré *et al*, 1996).

The sulfonation reaction results in the modification of both the N-termini and the ε-amino groups of lysine-containing peptides. For unprotected Lys-terminated peptides this will result in the formation of disulfonate derivatives. The formation of such derivatives was previously shown to be undesirable because they exhibit poor sensitivity in the positive-ion mode and relatively poor fragmentation under negative-ion analysis conditions. Negative-ion PSD spectra of several Lys-containing disulfonate derivatives showed both low-product ion yields and complex fragmentation patterns containing b- and y-type ions linked to the sulfonate group. Therefore, this approach requires a preliminary modification of the ε-amino group of lysine residues. It was demonstrated that, following guanidination of lysine ε-amines, the introduction of sulfonic acid groups to tryptic peptides is possible solely at the N-terminus. The guanidination, leading to lysines being converted into homoarginines (+ 42 Da), can selectively and quantitatively be performed with *O*-methylisourea at high pH and does not affect the peptide amino terminus or other side chains (Beardsley *et al*, 2002; Karty *et al*, 2002). Recently, we reported a novel approach in which gel-separated proteins are guanidinated in-gel prior to enzymatic cleavage. In contrast to previously described techniques, this procedure allows the extracted tryptic peptides to be N-terminal sulfonated without any further sample purification. We demonstrated that the obtained information can be used to identify proteins using the sequence similarity search algorithm FASTS (Sergeant *et al*, 2005).

The fully sequenced model plant *Arabidopis* is an excellent model system to study many basic plant processes but has its clear limitations. Many researchers investigating other plant proteomes have already experienced the lack of genomic resources as a bottleneck to identify proteins by MS (PMF). For example, proteomic studies on maize, an economically important organism, have been compromised due to the lack of database resources and the inability to use available database resources effectively (Porubleva *et al*, 2001). Plant scientists recognize the limitations of PMF-protein identification and realize the prospects of using sequence similarity methods to contribute to proteomics projects of non-sequenced plant species (van Wijk, 2001; Liska *et al*, 2004b). Therefore, a number of search algorithms have been developed to identify proteins by similarity to existing databases. Shevchenko *et al* developed a BLAST2-based search algorithm termed MS-BLAST for mass spectrometry driven BLAST searches (Shevchenko *et al*, 2001). Mackey *et al* described two sequence similarity search algorithms (FASTS and FASTF) that use multiple short peptide sequences to identify homologous sequences in protein or DNA databases. FASTS is designed to use *de novo* sequence data from organisms lacking comprehensive proteome sequence data (Mackey *et al*, 2002). MS-Homology is a database searching tool from the UCSF Mass Spectrometry Facility that performs homology-based searches (Clauser *et al*, 1999).

In the study presented here, *de novo* sequence was derived using our improved MS identification approach in combination with three different homology-based search algorithms. This approach allowed to identify a number of 2D-PAGE separated proteins from 2 banana varieties ITC 0084 and 0643 (*Musa* spp). Surprisingly, we observed that the

identification scores of some search algorithms vary, according to the position of the sequences in the query, most notably when applying the MS BLAST algorithm and, although to a lesser extent, using the FASTS algorithm. Bananas and plantains are important throughout the developing countries of the (sub)tropics both as a subsistence and an export crop. They are ranked as the fourth most important food crop of the world and are the most consumed fruit (Frison *et al*, 1998). *Musa* spp. have a basic chromosome number of 11 with an estimated size of 6 Mbp (Lysak *et al*, 1999). The varying genome combinations of the *Musa acuminata* A genome, alone or in combination with the *Musa balbisiana* B genome, are at the origin of the cultivated banana varieties. The cultivated varieties of banana consumed as fresh fruit have a AA or AAA genome type while the plantains have a AB, ABB, or AAB genome type (Swennen *et al*, 1994). To our knowledge, this is the first time that *de novo* derived sequence information, in combination with homology search algorithms, is successfully used to identify proteins from an eukaryotic organism with a poorly characterized genome. Consulting the Entrez genome project, the Genbank database contains at this moment 255 protein entries for the genus *Musa*, while e.g. 84168 entries are characterized for the model organism *Arabidopsis.*

Experimental Procedures

Protein extraction and 2DE

The selected varieties Cachaco ITC 0643 (ABB cooking banana) and Mbwazirume ITC 0084 (AAA highland banana) belong to different genomic groups. Multiple shoot meristem cultures were initiated as described by Strosse *et al* (Strosse *et al*, 2006) and subsequently maintained on the standard control medium (MS medium supplemented with BAP). Samples were extracted and separated according to Carpentier *et al*. (Carpentier *et al*, 2005). Banana meristems were excised and ground in a mortar in the presence of liquid nitrogen. 30-50 mg of protein was resuspended in 500 µl of ice-cold extraction buffer (50 mM Tris-HCl pH 8.5 ; 5 mM EDTA; 100 mM KCl; 1% (w/v) DTT; 30% (w/v) sucrose; complete protease inhibitor cocktail (Roche Applied Science, Vilvoorde, Belgium) and vortexed for 30 sec. 500 µl of ice-cold Tris buffered phenol (pH 8.0) was added and the sample was vortexed for 15 min at 4°C. After centrifugation (3 min, 6000g, 4°C) the phenolic phase was collected, re-extracted with 500 µl of extraction buffer and vortexed for 30 sec. After centrifugation (3 min, 6000g, 4°C) the phenolic phase was collected and precipitated overnight with 5 volumes 100 mM ammonium acetate in methanol at –20°C. The sample was diluted with rehydration buffer (6 M urea; 2 M thiourea; 0.5 % CHAPS; 10 % glycerol; 0.002% bromophenol blue; 0.5% IPG-buffer; 0.28% DTT) to 400 µg protein per 150 µl was applied via anodic cup loading. The concentration of the protein mixture was estimated using the 2-D quant kit from GE Healthcare (Diegem, Belgium). 24 cm IPG strips (GE Healthcare) were rehydrated for at least 8 h in 450 µl rehydration buffer. Isoelectric focusing was carried out on the IPGphor II (GE Healthcare) at 20 °C with a current limit of 50 µA/strip: 3h at 300V, 6h at 1000 V, 3h at 8000V (gradient) and 32000 Vh at 8000V. Prior to the second dimension the individual strips were equilibrated for 15 min in 6 ml equilibration solution (6 M urea; 30% glycerol; 2% SDS; 0.002% bromophenol blue; 50 mM Tris, pH 8.8) containing 1% (w/v) DTT and subsequently for 15 min in 6 ml equilibration buffer containing 4.5 % (w/v) iodoacetamide. The separation in the second dimension was performed in the Ettan DALT 6 System (GE Healthcare) with lab casted 1.5 mm SDS polyacrylamide gels (12.5%):45 min 12 W, 5 h 100 W. Acrylamide and protein standard were purchased from Biorad (Nazareth, Belgium). Proteins were visualized by colloidal Coomassie brilliant blue

staining (24). Gels were scanned and calibrated with labscan 5 software (GE Healthcare). Image analysis was performed with Image Master 2D platinum (GE Healthcare).

In-gel guanidination

Guanidination was performed by adding 5 µl MQ, 11 µl 7 N ammonium hydroxide (Merck, Darmstadt, Germany) and 3 µl of a 7.5 M *O*-methylisourea hemisulfate (50 mg in 51 µl MQ, prepared daily) (Across, Geel, Belgium) solution to the gel plugs. The samples were vortexed briefly and incubated at 65°C. After an incubation of two hours the guanidinated samples were taken from the oven and the remainder of the solution was discarded. The gel pieces containing the guanidinated samples were desalted and destained in one step. Two washes using 150 µl 200 mM ammonium bicarbonate in 50 % ACN/MQ (30 minutes at 30°C) were performed and, subsequently, the gel plugs were dried in a SpeedVac (Thermo Savant, Holbrook, NY).

Trypsin digestion and sulfonation

To the dried gel plugs, 8 µl digestion buffer (50 mM ammonium bicarbonate, pH 7.8) containing 150 ng modified trypsin per µl (Promega, Madison, WI) was added and the tubes were kept on ice for 45 minutes to allow the gel plugs to be completely soaked with trypsin. Digestion was performed overnight at 37°C, the supernatant was recovered and the resulting peptides extracted twice with 35 µl 60% ACN/0.1% DIEA. The extracts were pooled and dried in the SpeedVac. The peptides were redissolved in 4 µl 12.5 mM ammonium bicarbonate 50% ACN/MQ and 2 µl was mixed with 2 µl of the sulfonation solution. The sulfonation reagent was prepared by dissolving 2 mg 2-sulfobenzoic acid cyclic anhydride (Fluka, Buchs, Switzerland) in 1 ml dry tetrahydrofuran (THF) to attain a 0.01 mM solution. The tubes were briefly vortexed and allowed to react for 15 minutes at room temperature.

MS and MS/MS

A 4700 Proteomics Analyzer (Applied Biosystems, Foster City, CA) with TOF/TOF optics was used for all MALDI MS and MS/MS applications. This mass spectrometer uses a 200-Hz frequency tripled Nd:YAG laser at a wavelength of 355 nm. For MS/MS, ions generated by the MALDI process were accelerated at 8 kV through a grid at 6.7 kV into a short, linear, field-free drift region. In this region, the ions pass through a timed-ion-selector (TIS) device that is able to select a precursor for subsequent fragmentation in the collision cell. After a peptide at a given m/z was selected, it passed through a retarding lens where the ions were decelerated and then passed into the collision cell, which was operated at 7 kV. The collision energy is defined by the potential difference between the source and the collision cell (1 kV). After passing through the collision cell, the ions (both intact peptide and fragments) were accelerated in the second source region at 15 kV, passed through a second, field-free, linear drift region, into the reflector, and finally, to the detector. The detector amplifies and converts the signal to electric current, which is observed and manipulated by a PC-based operating system. For high resolution MS analysis, the instrument was operated in reflectron mode. After the MALDI process generates the peptide ions, the latter are accelerated at 20 kV through a grid at 14 kV into the first, short, linear, field-free drift region. After this point, the rest of the instrument can be treated as a continuation of this region until the ions enter the reflector and are reflected onto the detector.

Samples were prepared by mixing 0.7 µl of the sample with 0.7 µl matrix solution (7 mg/ml α-cyano-4-hydroxycinnamic acid (CHCA) in 50% ACN containing 0.1% TFA) and

spotted on a stainless steel 192-well target plate. They were allowed to air-dry at room temperature, and were then inserted in the mass spectrometer and subjected to mass spectrometric analysis. Prior to analysis, the spectrometer was externally calibrated with a mixture of Angiotensin I, Glu-fibrino-peptide B, ACTH (1-17), and ACTH (18-39). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrino-peptide B.

Database searches

The *de novo* determined peptide sequences were used for similarity searches using the FASTS, MS BLAST and MS-Homology algorithms. On-line submissions were performed using MS BLAST at the Heidelberg server (http://dove.embl-heidelberg.de/Blast2/msblast.html). Searches were performed against the non-redundant database (nrdb) using standard settings. The FASTS algorithm (http://fasta.bioch.virginia.edu/fasta_www/cgi/) was carried out using standard settings, and searches were performed against the NCBI/BLAST nrdb with BLOSUM 50 as search matrix. MS-Homology searches (Protein Prospector 4.0.5) were performed on the UCSF server against the NCBI nrdb using BLOSUM 50 as search matrix (http://prospector.ucsf.edu/ucsfhtml4.0/mshomology.htm).

The software used for similarity searches does not discriminate between the isobaric amino acids Ile and Leu. Therefore, all mass increments of 113 Da between consecutive y-ions were arbitrarily designated as Ile. The FASTS search results were considered significant if the E-value was below 1.0 e-04. The MS BLAST search results were considered significant if the resulting scores were higher than the threshold score indicated in the software. In order for a particular protein in the database to generate a hit, MS-Homology must find homologous sequences for the minimum number of peptides. The scoring method used is based on a mutation matrix, such as the one used in the BLAST and FASTA programs. The final score is calculated by adding the scores for the individual peptide alignments together. If there are several possible alignments of a given peptide, then the highest scoring alignment is used in the calculation. As the searches are based on similarity, proteins identified with lower scores must have the same generic function as the first hit. Proteins were considered as being positively identified only if all three search algorithms yielded the same homologous protein in the first hit. It has been demonstrated that indirect evidence can add to the significance of an identification (Shevchenko *et al*, 2001). Therefore, the identifications were further validated by using information such as the cleavage specificity of trypsin and sequence information resulting from known preferential fragmentation patterns of sulfonated peptides (Samyn *et al*, 2004).

Results

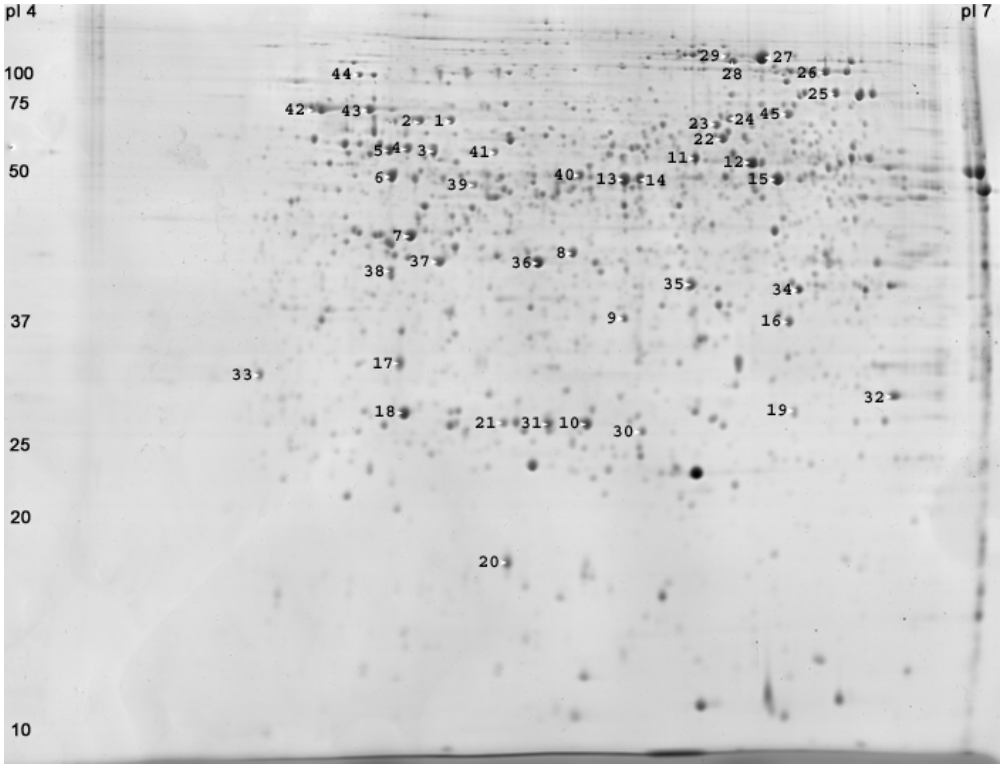


Figure 3.23. 2D-PAGE separated proteins from *Musa* variety ITC 0084. The analysed spots are numbered as in Table 1a. The horizontal bar indicates pI (IPG strip 4-7) and the vertical one Mw (kDa).

After Coomassie staining, 60 spots were randomly selected from both varieties (Figure 3.23). The proteins were guanidinated in-gel and desalted/destained in one single step as described previously. Subsequently, the guanidinated proteins were enzymatically cleaved with trypsin and, after extraction; the peptides were sulfonated (Sergeant *et al*, 2005). Upon derivatization, the peptides were subjected to MS/MS using a MALDI TOF/TOF instrument. Fragmentation of the derivatized peptides occurred under metastable decay conditions using a frame collision energy of 1 keV (no gas in the collision cell). In all experiments a CHCA matrix was used, as this is the most common matrix for peptide analysis. In all MS/MS analyses, a contiguous series of y-ions spectra was observed in the fragment spectra after guanidination and sulfonation. In the fragment spectra, an initial loss of the sulfonic acid derivative was observed (Δm = 184 Da). By simple manual calculation of the differences between the adjacent y-ion fragments the amino acid sequence could readily be interpreted. As observed before, in some experiments the sulfonic acid-derivatized peptides had poorer positive-ion sensitivity than the corresponding native peptides (Samyn *et al*, 2004). After sulfonation, some of the tryptic peptides were no longer observed in positive mode reflectron analysis. However, these fragments could be detected as their deprotonated ions when the analysis was performed in the negative mode. Selection of the corresponding protonated precursor ion (+2 Da) for MS/MS analysis (positive mode) resulted in the formation of a complete series of y fragment-ions. By way of example, we show in Figure 3.24 the derivatized tryptic fragments of the protein identified as F1-ATP synthase (Table 3.8a, spot 6, *Musa* variety ITC 0084). In the spectrum in positive reflectron analysis apparently only two peptides had a mass increase corresponding with the addition of a sulfonic acid group (+184 Da) (Figure 3.24a & b; m/z 1583.82 and 3113.61). However, MS analysis in negative reflectron mode indicated that multiple peptides are sulfonated (Figure 3.24a & c) (Δm = 184 - 2 Da). Figure 3.24d shows the fragmentation spectrum (positive mode) of the theoretical

precursor at m/z 1677.49 from which the complete sequence could be deduced. Most likely the protonated precursors are not detected due to metastable decomposition, but do yield excellent MS/MS spectra in the positive ion mode.

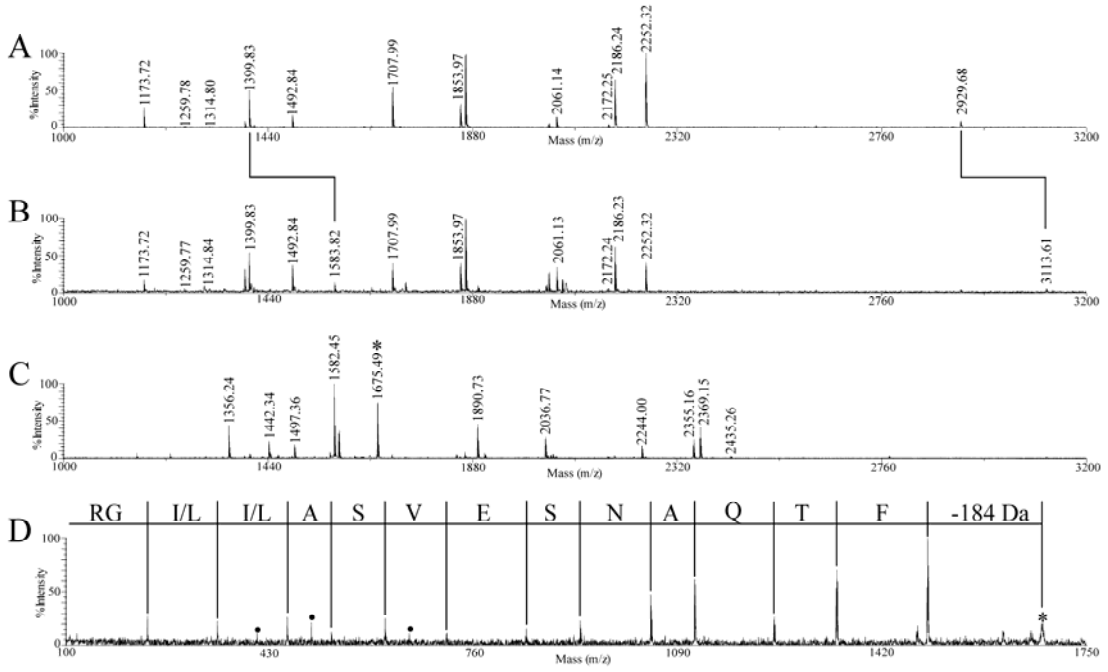


Figure 3.24. a) MALDI reflectron MS spectrum (positive mode) of the tryptic fraction after in-gel guanidination of spot 6 (*Musa* variety ITC 0084). Panels b and c respectively show the positive and negative MALDI MS spectrum after sulfonation (the mass increase of 184 Da, the sulfonation label, is indicated between panel a and b). d) MALDI MS/MS spectrum of the theoretical precursor at m/z 1677.49 (indicated with * in panel c, m/z 1675.49). The loss of the sulfonation label (-184 Da) is indicated. The *de novo* derived sequence information is given in the one-letter code.

The *de novo* derived (partial) sequence information from each spot was combined in one search query and analyzed using three search algorithms: MS BLAST, FASTS and MS-Homology (Shevchenko *et al*, 2001; Mackey *et al*, 2002). Most search queries included 20 to 60 amino acids, resulting from two to six peptide sequences (Appendix IIIa & b). Using these queries, all three homology-based search algorithms yielded identifications with a score significantly better than the threshold score (Table 3.8a & b). Only when the top results (first hits) from the three searches yielded the same protein, the identification was considered as positive. Using this approach, we were able to identify 40 out of 60 attempted spots from the 2D-PAGE maps, representing 31 unique proteins. In 5 spots, no tryptic peptides were observed upon in-gel guanidination. For 15 other spots, one or more peptides were observed after derivatization. Although MS/MS analysis yielded sequence information, these proteins were not identified when searching against the NCBI-database. This is most likely due to the fact that insufficient sequence information was obtained (e.g. only one peptide) or because of the lack of sufficiently homologous proteins in the database. When trying to identify proteins by sequence similarity searches, the number of peptides recognized from a digested protein determines the success of the identification. It has been calculated that as more peptides are analyzed and matched, proteins of less similarity to database sequences can be identified with the limit being around 50% identity (Mackey *et al*, 2002). For the 40 identified proteins, the majority of cross-species hits were made to proteins from plants with (completely) sequenced genomes such as rice (*Oryza sativa*) or *Arabidopsis* (Table 3.9).

Table 3.8a. 2D-PAGE separated proteins from *Musa* variety ITC 0084

Spot[a]	Protein[b]	Identification FASTS[c]	Pept.[d]	AA[e]	E-score[f]	MS BLAST score[g]	MS-Hom. score[h]
1-2	7330642	HSP68	9	78	3.2e-25	277	223
3	52353541	Put. ketol-acid reductoisomerase	5	40	4.9e-18	183	226
4	12546	Chaperonin 60	5	46	4.2e-19	223	244
5	31432537	Mitochondrial chaperonin 60	4	27	6.8e-08	121	124
6	4388534	F1-ATP synthase	8	100	3.4e-53	579	427
7	3746942	Actin 1	4	46	1.7e-23	215	270
8	52076544	Put. Cytoxolic phosphogly. Kin.	6	42	5.3e-08	141	162
9	57014097	Pectinesterase 3 precursor	3	24	2.7e-12	153	163
10	-	-	2	14	-	-	-
11/12	10716961	Polyophenol oxidase	2	16	2.1e-06	111	112
11	-	-	3	37	-	-	-
12	18479040	26S proteasome reg. subunit IV	2	17	1.8-09	124	117
12	12802327	Mito. proc. peptidase β-subunit	1	21	8.9e-10	108	113
13	33113259	Enolase	3	34	3.8e-13	210	181
14	6601496	S-adenosylhomocyst. hydrolase	6	53	1.4e-32	312	265
15	114411	ATP synthase α-chain	6	56	1.6e-27	304	282
16	-	-	3	16	-	-	-
17	-	-	3	32	-	-	-
18	37020723	Ascorbate peroxidase	4	51	5.7e-26	266	282
19	-	-	-	-	-	-	-
20	-	-	2	28	-	-	-
21	37783265	Ascorbate peroxidase	2	26	2.0e-07	92	97
22	-	-	1	8	-	-	-
23	26453278	Put. succ. dehydr. flavoprotein	3	34	1.5e-16	172	191
24	-	-	2	16	-	-	-
25	-	-	3	18	-	-	-
26	-	-	3	20	-	-	-
27	2369714	Elongation factor 2’	4	36	9.8e-25	271	249
28	-	-	1	10	-	-	-
29	50909007	Putative elongation factor 2	3	21	5.4e-06	111	108
30	-	-	-	-	-	-	-
31	27650423	Ascorbate peroxidase	3	27	7.0e-11	139	158
31	47607439	Mit. ATP synthase precursor	2	22	4.4e-04	104	107
32	4336905	Ran-related GTP binding protein	3	28	6.9e-17	184	199
33	7435012	14-3-3-protein tf6	3	37	5.3e-26	258	216
34	1658313	Osr40g2	4	40	3.0e-10	145	205
35	2286153	Cytosolic malate dehydrogenase	2	24	6.3e-14	137	113
36	1527223	Glutamine synthetase	1	16	2.2e-12	126	111
37	50932771	Putative malate dehydrogenase	3	26	6.2e-11	132	135
38	-	-	2	12	-	-	-
39	6136112	UTP-gluc-1-phos. uridyltransf.	1	12	1.4e-06	94	86
40	33113259	Enolase	2	19	4.8e-09	130	111
41	4206124	T-complex protein 1 ε-subunit	2	21	5.5e-13	151	146
42-43	56554972	Heat shock protein 70	3	31	2.7e-16	186	128
44	-	-	1	7	-	-	
45	-	-	2	14	-	-	

Table 3.8b. 2D-PAGE separated proteins from *Musa* variety ITC 0643

Spot[a]	Protein[b]	Identification FASTS[c]	Pept.[d]	AA[e]	E-score[f]	MS-BLAST score[g]	MS-Hom. score[h]
1	-	-	-	-	-	-	-
2	-	-	1	8	-	-	-
3	41818408	Class III acidic chitinase	3	26	1.3e-09	163	148
4	-	-	-	-	-	-	
5	39939493	Ascorbate peroxidase	6	61	7.3e-16	203	211
6	1296955	r40c1 protein	8	63	2.6e-17	226	209
7	37928995	Cytosolic malate dehydrogenase	4	30	2.3e-07	171	120
8	56202334	Alpha-amylase isozyme III	4	31	3.2e-09	72	191
9	-	-	1	9	-	-	
10	25809056	DEAD box RNA helicase	2	16	2.9e-07	84	105

11	-	-	1	6	-	-	
12	-	-	-	-	-	-	
13	-	-	3	24	-	-	
14	55297085	Put. ketol-acid reductoisomerase	2	23	3.7e-07	80	105
15	-	-	1	5	-	-	

[a] Spot number according to the position on the 2D-PAGE (Figure 1)
[b] NCBI *Entrez* entries (http://www.ncbi.nih.gov/Entrez/)
[c] Identification based on FASTS search result
[d] Number of peptide sequences used in the query
[e] Total number of amino acids used in the query
[f] In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI non-redundant protein databases N ≈ 2075116.
[g] MS BLAST score for searches against the non-redundant protein database at http://dove.embl-heidelberg.de/Blast2/msblast.html
[h] MS-Homology (Protein Prospector 4.0.5) score against the NCBI non-redundant protein database.

FASTS searches databases using peptide sequences of unknown order, thereby evaluating all possible arrangements of the peptides. Because the true order of the query peptides used by FASTS is not known, FASTS only requires that the aligned peptides do not overlap. The algorithm is based on the heuristic FASTA comparison strategy to accelerate the search but uses alignment probability, rather than a similarity score, as the criterion for alignment optimality (Mackey *et al*, 2002). MS BLAST utilizes redundant, degenerate, and partially inaccurate peptide sequence data as is obtained by automated the *de novo* interpretation of MS/MS spectra. MS BLAST does not allow gaps within individual peptides, while gaps between peptides are not penalized and can be of arbitrary length. Therefore, all peptide sequences obtained by the interpretation of acquired MS/MS are assembled into a single searching string in arbitrary order (Shevchenko *et al*, 2001). MS-Homology (Protein Prospector 4.0.5) is a program that allows comparing a number of *de novo* derived peptide sequences, followed by the maximum number of amino acid substitutions allowed for each sequence, against a selected database (Clauser *et al*, 1999). Different peptides from the same unknown protein can be entered in the list. A database search will look for proteins containing peptides identical or homologous to the listed sequences. The quality of the results will be dependent on the number of peptides sequenced and the accuracy of the sequence information entered, as well as on database completeness and species to species sequence variability for the peptides entered. It is also possible to enter a part of the sequence as a mass, along with a tolerance factor.

Table 3.9. FASTS search result for spot 6 (Musa variety ITC 0084)

Hit	Protein[a]	Identification	Species	E-Score[b]	#pept[c]
1	4388533	F1-ATP synth, beta subunit	*Sorghum bicolor*	3.4e-53	7/8
2	50932681	putative ATP synth beta chain	*Oryza sativa*	1.5e-52	7/8
3	22173	unnamed protein product	*Zea mays*	1.6e-52	7/8
4	34911264	putative ATP synth beta chain	*Oryza sativa*	1.6e-52	7/8
5	3893824	ATPase beta subunit	*Nicotiana sylvestris*	6.5e-51	7/8
6	3893822	ATPase beta subunit	*Nicotiana sylvestris*	6.6e-51	7/8
7	3676296	mito ATPase beta subunit	*Nicotiana sylvestris*	6.7e-51	7/8
8	19685	ATP synthase beta subunit	*Nicotiana plumbaginifolia*	7.1e-51	7/8
9	3676294	mito ATPase beta subunit	*Nicotiana sylvestris*	7.2e-51	7/8
10	525291	ATP synth beta subunit	*Triticum aestivum*	1.3e-49	7/8
11	231587	ATP synth beta chain	*Oryza sativa*	4.5e-48	7/8
12	4388534	F1-ATP synth, beta subunit	*Sorghum bicolor*	2.2e-47	6/8
13	18831	mito ATP synth beta-subunit	*Hevea brasiliensis*	2.6e-45	7/8
14	56784992	put ATP synth beta subunit	*Oryza sativa*	1.6e-44	6/8
15	56784991	put ATP synth beta subunit	*Oryza sativa*	1.8e-44	6/8
16	23397307	unknown protein	*Arabidopsis thaliana*	9.9e-43	7/8
17	18415909	ATP bind. / H-exp ATPase, phosphorylative mechanism	*Arabidopsis thaliana*	5.5e-42	7/8
18	18415911	ATP bind/ H-exp ATPase, phosphorylative mechanism	*Arabidopsis thaliana*	5.5e-42	7/8
19	22326673	ATP bind / H-exp ATPase, phosphorylative mechanism	*Arabidopsis thaliana*	5.7e-42	7/8
20	17939849	Mit. F1 ATP synth beta subu.	*Arabidopsis thaliana*	8.5e-42	7/8
21	18322	ATP synth b subunit	*Daucus carota*	6.2e-37	6/8
22	2116558	F1 ATPase	*Pisum sativum*	2.7e-30	6/8
23	77546804	ATP synthase F1, b subunit	*Pelobacter carbinolicus*	4.4e-30	7/8
24	77544640	ATP synthase F1, b subunit	*Pelobacter carbinolicus*	4.4e-30	7/8
…					
2728	9909941	ATP synth beta subunit	*Rhizobium tropici*	8.8e-05	2/8
2729	62637659	ATP synth beta subunit	*Rhizobium gallicum*	8.8e-05	2/8
2730	9909586	ATP synth beta subunit	*Agrobacterium rhizogenes*	8.8e-05	2/8
2731	34499998	ATP synth beta subunit	*Mannheimia*	9.6e-05	2/8
2732	34500000	ATP synth beta subunit	*Mannheimia haemolytica*	9.6e-05	3/8
2733	34500010	ATP synth beta subunit	*Haemophilus parasuis*	9.6e-05	3/8

[a] NCBI *Entrez* entries (http://www.ncbi.nih.gov/Entrez/)
[b] In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI non-redundant protein databases N ≈ 2075116.
[c] number of matched peptides in the query

Surprisingly, we observed that the identification scores obtained varied according to the position of the peptide sequences in the query using MS BLAST and, to a lesser extent, using the FASTS search algorithm. As an example, the results from the three search algorithms are listed in Table 3.10. In spot 29 three peptide sequences were determined *de novo* (Appendix IIIa). All six possible variations of the positions of the peptide sequences in the query were submitted to the three search algorithms. The MS-Homology algorithm yielded the same protein for all queries with an identical, significant, score. The same result was observed using the FASTS algorithm, although we observed a minor variation in the search score. From the latter results it appeared that the shortest peptide (VIKI) was not used for identification in all possible queries. The same protein, putative elongation factor 2, was identified using MS BLAST. However, in two queries the search score dropped below the

significance level, most likely because only one peptide sequence, AMKFSVSP, was matched to the protein. In one query, where the latter sequence was not matched to the protein, the MS BLAST algorithm identified another protein, a hypothetical yeast protein (Q6CWD9) with a score below the significance level (Table 3.10). Ungapped BLAST search identifies all high-scoring pairs (HSP), regions of high local sequence similarity between individual peptides in the query and a protein sequence from a database entry. It has been suggested that the sequential order of the matched segments does not affect the total score, which is calculated for each protein entry by adding up the scores of individual HSPs that are higher than the specified threshold (Shevchenko *et al*, 2001). This is confirmed by our results. Using the three first MS BLAST queries, e.g., results in the matching of the same two peptides and, hence, the same score (Table 3.10). However, as indicated by our results, merging the peptides in another, non-arbitrary order, results in identification with a lower, non-significant, identification score or even another protein.

Table 3.10. Search results for spot 29 (*Musa* variety ITC 0084)

E-score[a]	Order peptide sequences[d]	MS BLAST score[b]	Order peptide sequences[d]	MS-Homol. score[c]	Order peptide sequences[d]
5.3e-06	VKFTXXEIR VIKI AMKFSVSP	111	VKFTXXEIR VIKI AMKFSVSP	108	VKFTXXEIR VIKI AMKFSVSP
5.1e-06	VKFTXXEIR AMKFSVSP VIKI	111	VKFTXXEIR AMKFSVSP VIKI	108	VKFTXXEIR AMKFSVSP VIKI
5.4e-06	VIKI VKFTXXEIR AMKFSVSP	111	VIKI VKFTXXEIR AMKFSVSP	108	VIKI VKFTXXEIR AMKFSVSP
5.4e-06	VIKI AMKFSVSP VKFTXXEIR	60[e]	VIKI AMKFSVSP VKFTXXEIR	108	VIKI AMKFSVSP VKFTXXEIR
5.3e-06	AMKFSVSP VIKI VKFTXXEIR	60[e]	AMKFSVSP VIKI VKFTXXEIR	108	AMKFSVSP VIKI VKFTXXEIR
5.1e-06	AMKFSVSP VKFTXXEIR VIKI	84[e]	AMKFSVSP VKFTXXEIR VIKI	108	AMKFSVSP VKFTXXEIR VIKI

[a] In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI non-redundant protein databases (N ≈ 2075116).
[b] MS BLAST score for searches against the non-redundant protein database at http://dove.embl-heidelberg.de/Blast2/msblast.html.
[c] MS-Homology (Protein Prospector 4.0.5) score against the NCBI non-redundant protein database.
[d] Order of the peptide sequences used in the different queries. Peptide sequences matched against the identified peptide are indicate in red.
[e] MS BLAST scores below treshold score (not significant).

In contrast to organisms in which one gene gives rise to a single protein (many of the viral, archaea and prokaryotic proteins), higher eukaryotes like plants tend to have more, but very similar proteins. Isoforms exist by multiple mechanisms: different gene loci, multiple

alleles, different subunit interactions, different splice forms, or different post-translational modifications. Isoforms are usually separated during two-dimensional gel electrophoresis. In spot 9 from *Musa* variety ITC 0084, we identified two isoforms from the pectinesterase 3 precursor (Table 3.8a). The MS/MS spectra of the derivatized peptides at m/z 1608.84 and m/z 1574.72 yielded the sequences SATFAVVGE and SAT(I/L)AVVGEGF(I/L)AR respectively (Figure 3.25). The observed mass difference (34.1 Da) between the two peptides fits exactly with the mass difference between Phe (147.2) and Leu/Ile (113.2) at position 4 in the sequence, whereas the remainder of the peptide sequence is identical. Variety ITC 0084 is a triploid AAA. The triplets coding for Phe (UUU, UUC) and Leu (UUG, UUA, CUU, CUC, CUG and CUA) are very closely related and, therefore, both isoforms are probably the result of a single nucleotide polymorphism (SNP) event, representing an allelic variance.

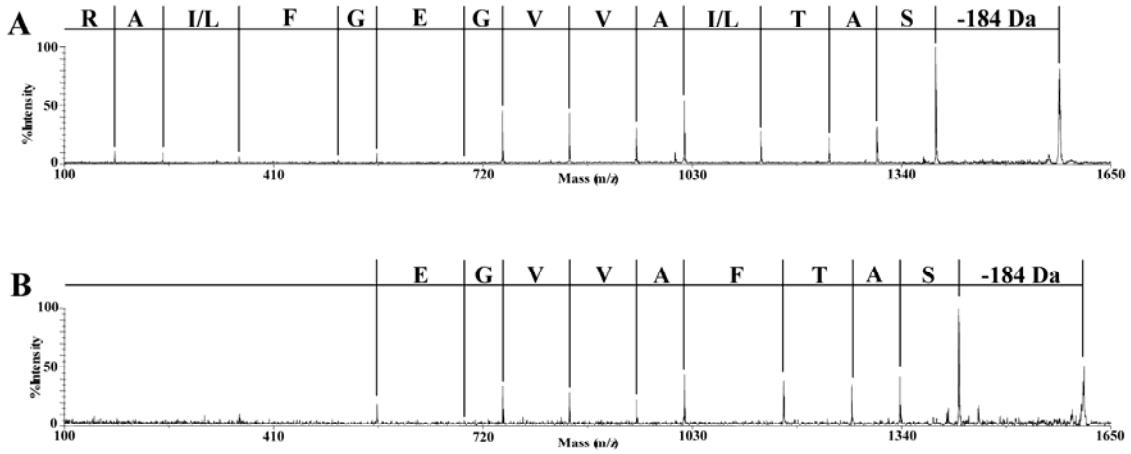


Figure 3.25. Identification of two isoforms from the pectinesterase 3 precursor in spot 9 from *Musa* variety ITC 0084. MALDI MS/MS spectrum (positive ion mode) of precursor at m/z 1574.72 (a) and m/z 1608.84 (b).). All labeled fragment ions are y-ions; the loss of the sulfonation label is indicated as a loss of -184 Da. The *de novo* derived sequence information is given in the one-letter code.

Similarly, in spot 5 from *Musa* variety ITC 0643 we observed two ascorbate peroxidase isoforms with an estimated Mw of 27 kDa and a pI of 5.5. MALDI TOF/TOF analysis of the peptides at m/z 2048.96 and m/z 2021.96 yielded the sequences FPAEIAHGADDGINI and FPAEIAHGADDGISI respectively, differing in their penultimate amino acid (Results not shown). Although the codons for Asn and Ser are not related, excluding the possibility of a SNP, we anticipate that the presence of an Asn is correct as the mass difference of 27 Da cannot be explained by any known post-translational modification of the penultimate Ser. Variety ITC 0643 is an ABB triploid, originating from the two different wild type genomes A and B. So, most likely, the isoforms result from an allelic variance, originating from the two genetically different wild type alleles. The sequence FPAEIAHGADDGINI was also observed in spot 18 from variety ITC 0084, a triploid AAA. Furthermore, ascorbate peroxidase was also identified in other spots from this variety (spot 18, 21 & 31 Table 3.8a). The proteins separate in 2D-PAGE as 3 individual spots with an estimated Mw of 27 kDa and a pI of respectively 5.50, 5.75 and 5.94. Alignment of the peptide sequences (Figure 3.26) indicates that the isoforms do not originate from a posttranslational modification but are probably different gene loci. A search on the Entrez *Arabidopsis* genome project reveals 7 different gene loci for L-ascorbate peroxidase.

Spot 18	F	P	**A**	**E**	**I**	**A**	**H**	**G**	**A**	**D**	**D**	**G**	**I**	N	**I**		
Spot 21	**I**	**E**	**A**	**E**	S	**A**	**H**	**G**	**A**	N	**D**	**G**	**I**	**D**	**I**	A	V	R
Spot 31	**I**	**E**	**A**	**E**	**I**	**A**	**H**	**G**	**A**	**D**	**D**	**G**	**I**	**D**	**I**			

Figure 3.26. Sequence alignment of peptides observed in spots containing ascorbate peroxidase. (*Musa* variety ITC 0084)

Another example of a variety specific isoform was observed in spot 3 (variety ITC 0084) and 14 (variety ITC 0643). Both were identified as a ketol-acid reductoisomerase. Both proteins have an estimated Mw of 61 kDa but spot 14 has a pI of 5.59 whereas spot 3 has a pI of 6.00. The difference in pI is related to different sequences. The peptide GVAYMV has been identified in variety ITC 0084**,** while GVSFMV was identified in variety ITC 0643.

In spots 4 and 5 from *Musa* variety ITC 0084 the protein was identified as a chaperonin 60 protein. Both spots have an estimated Mw of 63 kDa and a pI of respectively 5.53 and 5.47. In spot 4 we observed the sequence GITMAVDSVVTN (MS/MS of m/z 1900.98) identifying the protein as chaperonin 60 from a *Cucurbita* species whereas MS/MS analysis of spot 5 yielded the sequence GISMAVDSVVTN (m/z 1886.98), identifying the protein as a mitochondrial chaperonin 60 from rice (Table 3.9a). Alignment of both proteins gave a sequence identity of 88%. Furthermore, both MS/MS spectra show an initial loss of 201 Da rather than the expected loss of the sulfonation label (-184 Da) (Figure 3.27). The y-ion series is evident in the spectra, but all y-ions are accompanied by (y-17)-ions which are approximately twice as large as the corresponding y-ions. The formation of this second series is due to the presence of an internal homoarginine inducing a neutral loss of ammonia (-17 Da). The loss of such neutral molecules has also been observed during MALDI-analysis of peptides containing internal Arg residues (Samyn *et al*, 2004).

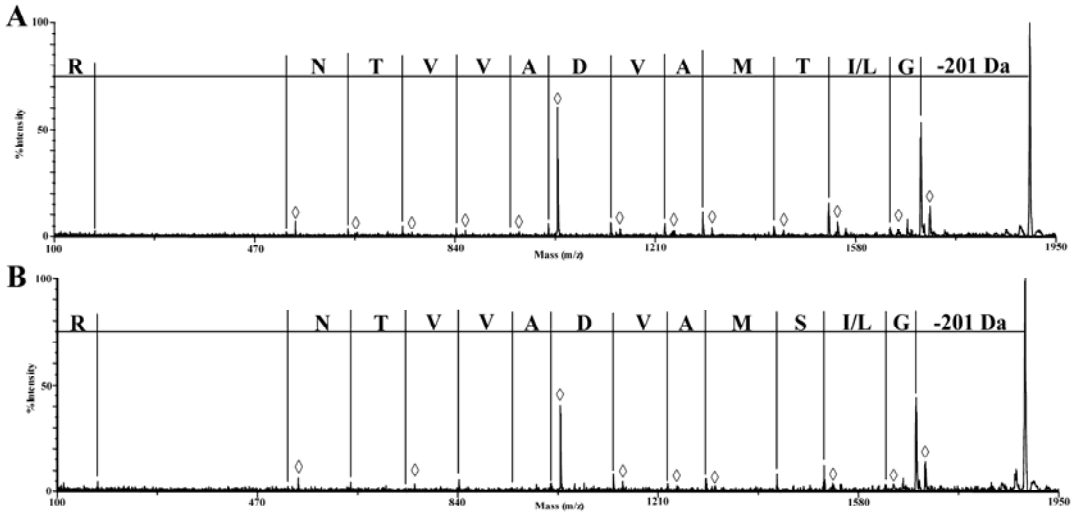


Figure 3.27. MALDI MS/MS spectrum (positive ion mode) of a derivatized peptide isolated from spot 4 (a) (precursor m/z 1900.98) and of a derivatized peptide from spot 5 (b) (precursor m/z 1886.98). All labeled fragment ions are (y-17)-ions resulting from the neutral loss of NH3; y-ions are indicated as ◊. The de novo derived sequence information is indicated in the one-letter code. The loss of the sulfonation label (-184 Da) is indicated as -201 Da, including the neutral loss of ammonia (-17 Da).

Discussion

Plant proteomics so far has encompassed a variety of species, including *Arabidopsis* (*Arabidopsis thaliana)* (Mayfield *et al*, 2001), rice (*Oryza sativa)* (Koller *et al*, 2002), maize (*Zea mays*) (Chang *et al*, 2000), pea (*Pisum sativum*) (Peltier *et al*, 2000) and wheat (*Triticum aestivum*) (Amiour *et al*, 2002). Recently, a number of genome and proteome studies have been performed to study stress responses in plants. However, most of them were performed on model plant organisms with completed genomes such as *Arabidopsis* or rice (Boudart *et al*, 2005; Denby *et al*, 2005; Yan *et al*, 2005; Agrawal *et al*, 2006; Ali *et al*, 2006).

For organisms whose genomes are not or only partially known, identification depends on non-error-tolerant MS/MS database searching or the more sensitive, *de novo* sequence similarity database searching, as recently reviewed by Liska and Shevchenko (Liska *et al*, 2003b). MS BLAST has successfully been applied for the identification of unknown proteins from the Brazilian moth, *Cerodirphia speciosa* (Shevchenko *et al*, 2005), the African clawed frog *Xenopus laevis* (Liska *et al*, 2004a), Dead Sea alga *Dunaliella salina* (Liska *et al*, 2004b), methylotrophic yeast *Pichia pastoris* (Shevchenko *et al*, 2001), and holm oak leafs (Jorge *et al*, 2005). Wait *et al* used BLAST and FASTS to identify proteins isolated from bovine serum if no identification resulted from PMF-analysis or MASCOT searches against a dbEST database (Wait *et al*, 2002). However, only ten percent of the proteins were identified by similarity searches to a non-bovine protein homologue. More recently, Matis *et al* used a combination of isotopic labeling with MS/MS for *de novo* sequence analysis of proteins from the fungus *Pleurotus ostreatus*. Due to the complexity of metabolism and morphology of the fungi, there is a strong reduction in similarity to other organisms at the genomic and proteomic level. Using FASTS they identified 4 proteins, in an SDS-PAGE separated fraction, which shared as low as 60% homology with proteins in databases (Matis *et al*, 2005).

Here, we demonstrated that sequence similarity searches substantially expand the boundaries of proteomics in plants whose genomes are not known. Out of 60 attempted spots, 40 (67 %) were identified, representing 31 unique proteins (Table 3.9a & b). As was anticipated, all identifications were produced by cross-species matching to known relatively conserved proteins from other plants (Table 3.10). When investigating the proteome of an organism with an unsequenced genome, the ability to identify proteins is dependent on the content of available databases. Where an abundance of database sequences exists of closely related organisms, with respect to the organism under inquiry, more homologous genes exist in silico to make cross-species identifications possible. If the organism being studied is more distantly related to any organism with a sequenced genome, the likelihood of protein identification decreases (Liska *et al*, 2003a). The ongoing sequencing of plant genomes and ESTs both contribute to the increased representation of protein sequences in databases and will enable the characterization of proteins from more phylogenetically distant species. It is very likely that an additional number of the *Musa* proteins can be identified by performing the same search against an EST-database (Liska *et al*, 2003b). MS-Homology, MS BLAST and FASTS methods provide independent means of evaluating the statistical significance of hits and, therefore, it is not necessary to compare retrospectively the matched peptide sequences with actual tandem mass spectra to rule out false positive hits. As reported previously, the position of the *de novo* determined peptide sequences in the query were arbitrarily chosen (Shevchenko *et al*, 2001; Mackey *et al*, 2002). However, in this study we observed that the identification scores vary, according to the position of the sequences in the query, when applying the MS BLAST algorithm. As demonstrated, the identification score decreases below the threshold value, and in one particular query another protein was identified,

however, also with an identification score below the threshold value (Figure 3.26). This problem will be the subject of further study.

Although *de novo* sequencing of underivatized peptides using MALDI TOF/TOF has recently been demonstrated, the interpretation of fragment spectra from peptides originating from unknown proteins strongly depends on the use of automated search routines or on manual interpretation. TOF/TOF fragmentation analysis of underivatized peptides yields multiple, incomplete fragment ion series, which are often difficult to interpret. The introduction of a sulfo group facilitates the MS/MS fragmentation of singly charged peptide ions by providing a second, 'mobile' proton, which lowers amide bond strength and allows more facile unimolecular decay. N-terminal tags containing a sulfo group have been advocated as a useful approach to generate a full y-ion series of peptide fragments in MS/MS (Keough *et al*, 1999; Samyn *et al*, 2004; Sergeant *et al*, 2005). The *de novo* determined sequences from derivatized peptides normally yield a contiguous sequence between 5 and 20 amino acids. Here, we demonstrated that this information is sufficient to identify different isoforms resulting from SNPs, allelic variations, or from different gene loci.

With sequence-similarity database searching methods, the proteomes of plants with unsequenced genomes will be more amenable for characterization by high-throughput MS techniques. It enables the identification of more conserved proteins in species that are distantly related to plants with sequenced genomes, as well as more divers homologous proteins. Previous studies have indicated that sequence-similarity protein identification by MS can identify more proteins than conventional approaches (Liska *et al*, 2003b; Liska *et al*, 2003a; Liska *et al*, 2004a; Liska *et al*, 2004b).

3.3.3. Automation

The improved protocol for N-terminal sulfonation of peptides, including in-gel guanidination as described in Part 3.2, opens the possibility for automation of this approach. Fully automated trypsin digestions, starting from the excision of gel spots to the preparation of samples for analysis, can now routinely be performed.

3.3.3.1. SPITC, an alternative sulfonation reagent

Introduction

In previous studies, we demonstrated that the use of 2-sulfobenzoic acid cyclic anhydride (SACA) results in fast derivatization of peptides. As this reagent reacts with water, it must be dissolved in THF. However, this solvent poses problems for use in automated platforms. Therefore, a water compatible reagent for N-terminal sulfonation was tested prior to attempting automation and the results of both derivatization reagents have been compared. 4-sulfophenyl-isothiocyanate (SPITC) is a cheap, commercially available sulfonation reagent that is compatible with the use of aqueous solutions (Gevaert *et al*, 2001). SPITC has recently been applied in several studies as an alternative to the expensive 'chemically assisted fragmentation'-reagent (CAF) from GE Healthcare (Marekov *et al*, 2003; Lee *et al*, 2004a).

Here, we demonstrate that SPITC is a better alternative for N-terminal sulfonation then 2-sulfobenzoic acid cyclic anhydride. Complete derivatization was observed in most experiments, fragmentation spectra are simple and can be readily interpreted and, most importantly, we observed that the use of SPITC results in an increased sensitivity.

Materials and methods

Gel electrophoresis

The total protein extract from anaerobically grown *Shewanella oneidensis* (± 50 µg of protein as determined by a Bradford test) was electrophoretically separated according to Laemmli. 12% Tris-glycine gels with a thickness of 1 mm containing 10 wells were casted. Electrophoresis was carried out using a Mini-Protean III system (BioRad, Nazareth, Belgium). Protein samples were mixed 1:1 (v/v) with sample buffer containing β-mercaptoethanol as the reducing agent and bromophenol blue to visualize the electrophoresis front. The sample was briefly heated (90-95°C, 5 min) before it was loaded on the gel. The electrophoresis running buffer was 25 mM Tris base, 192 mM glycine, and 0.1% SDS (w/v). Electrophoresis was carried out at 150 V for +/- 1.5 hours, until the dye marker had reached the edge of the gel. After fixation (2% H_3PO_4/50% ethanol/MQ; 30 minutes), proteins were stained with CBB G-250 at 0.2% (w/v) in 34% methanol/17% ammonium sulfate, containing 3% phosphoric acid, for +/- 30 minutes. Background destaining was carried out overnight with a 30% methanol solution.

Guanidination

In-gel guanidination was performed according to a previously published protocol (Sergeant *et al*, 2005). Briefly, 5 µl MQ, 11 µl 7 N ammonium hydroxide (Merck, Darmstadt, Germany) and 3 µl of a 7.5 M *O*-methylisourea hemisulfate (Across, Geel, Belgium) solution were added to the excized SDS-PAGE bands. The samples were briefly vortexed and

incubated at 65°C. After an incubation of two hours the remainder of the reaction mixture was discarded. The gel pieces were desalted and destained in one step by two washes with 150 µl 200 mM NH_4HCO_3 in 50% ACN/water (30 min at 30°C) and dried in the SpeedVac (Thermo Savant, Holbrook, NY).

Tryptic digestions

A volume of 8 µl digestion buffer (50 mM ammonium bicarbonate, pH 7.8) containing 150 ng modified trypsin per µl (Promega, Madison, WI) was added to the dried gel spots. The tubes were kept on ice for 45 minutes to allow the gel pieces to be completely soaked with the protease solution. Digestion was performed overnight at 37°C, the supernatant was recovered and the resulting peptides extracted twice with 35 µl 60% ACN/0.1% DIEA. The extracts were pooled, divided in two vials and dried.

Sulfonation

For sulfonation using SACA the peptides were redissolved in 4 µl 12.5 mM NH_4HCO_3 50% ACN/MQ, and 0.7 µl was spotted and analyzed. 2 µl of the remainder was mixed with 2 µl of the sulfonation solution, prepared by dissolving 2 mg SACA in 1 ml dry THF (0.01 M) solution. The tubes were briefly vortexed and reacted for 15 minutes at room temperature. For SPITC-derivatization 4 µl 20 mM NH_4HCO_3 was added to the dried sample. 0.7 µl was used for spotting and to 2 µl of the remainder an equal volume of the SPITC-solution was added. SPITC-solution was prepared by dissolving 10 mg SPITC sodium salt in 1 ml 20 mM NH_4HCO_3 (40 mM). After briefly vortexing the sample, it was incubated at 37°C for 30 minutes to attain complete derivatization.

Mass spectrometry

The Applied Biosystems 4700 Proteomics Analyzer with TOF/TOF optics (Applied Biosystems, Foster City, CA) was used in this study for MALDI MS and MS/MS applications. This mass spectrometer uses a 200-Hz tripled Nd:YAG laser operating at a wavelength of 355 nm. For MS/MS, ions generated by the MALDI process were accelerated at 8 kV through a grid at 6.7 kV into a short, linear, field-free drift region. In this region, the ions passed through a timed-ion-selector device that is able to select one precursor for subsequent fragmentation in the collision cell. After selection the ions are decelerated and passed into the collision cell, which was operated at 7 kV. The collision energy is defined by the potential difference between the source and the collision cell (1 kV). After passing through the collision cell, the ions (both intact peptides and fragments) were accelerated in the second source region at 15 kV, passed through a second, field-free, linear drift region, into the reflector, and finally, to the detector. The detector amplifies and converts the signal to electric current, which is observed and manipulated with the software of the instrument. For high resolution MS analysis, the instrument was operated in reflectron mode. After the MALDI process generates the peptide ions, the latter are accelerated at 20 kV through a grid at 14 kV into the first, short, linear, field-free drift region. After this point, the rest of the instrument can be treated as a continuation of this region until the ions enter the reflector and finally reach the detector.

0.7 µl of sample was applied on the stainless steel probe and, after 5 minutes, when the volume was reduced about 50%, 0.5 µl matrix solution was added. Matrix was prepared fresh every day by dissolving 5 mg of 4-hydroxy-α-cyanocinnamic acid in 700 µl 50% ACN/MQ

0.1% TFA. The spot was allowed to dry at room temperature. MS analysis, before and after sulfonation, indicates the peptides that are sulfonated by a mass shift of 184 or 215 Da for peptides sulfonated respectively with SACA and SPITC. From previous studies (Samyn *et al*, 2004; Sergeant *et al*, 2005) it was known that attachment of a sulfonic acid moiety may result in the suppression of the peptide signal in positive mode MS. Nonetheless, these sulfonated peptides could be identified either in negative mode MS analysis or by simply adding the respective mass to peptide masses observed before sulfonation. For MS/MS, sulfonated peptides were selected allowing a -2 to +4 mass range around the precursor mass. 5000 shots were acquired for each precursor using 1kV fragmentation energy settings with no gas in the collision cell. All fragmentation spectra were interpreted manually.

Database searches

All sequences acquired from a single spot were initially submitted in a single database search. If not all sequences in a query matched to a protein, those peptides associated with an unambiguous match were omitted and the remaining sequences submitted in subsequent searches. For the identification of proteins with the *de novo* determined sequences three different search-algorithms were used. MS BLAST, FASTS and MS-Homology, all three were used as described before (Part 3.3.2). Hits were manually validated by taking the specificity of trypsin into account and by controlling the fragment ion spectra based on the known preferential fragmentation of certain peptide bonds (Samyn *et al*, 2004).

Results

50 µg of total cellular protein extract from *Shewanella oneidensis* was separated by SDS-PAGE. After staining, the bands indicated in Figure 3.28a were excized. Initially, 15 bands were picked and processed according to the protocol for in-gel guanidination. To ensure that exactly the same sample was used for both derivatization reagents, the sample was divided in two equal parts after trypsin digestion and extraction of the peptides. The aliquots were dried and processed according to our protocol for derivatization with SACA (Sergeant *et al*, 2005) or according to a modified protocol for SPITC-derivatization (Wang *et al*, 2004). Analysis of guanidinated samples, after drying of the extract and redissolving in 50% ACN, prior to sulfonation revealed that few or no peaks were observed in the spectra that were prepared for sulfonation with SACA. On the contrary, in MS spectra from samples prepared for sulfonation with SPITC, reconstituted in water, numerous peptides were observed. After performing the respective sulfonation protocols, a similar trend was observed for the sulfonated samples, both in positive and negative mode.

From the MS-spectra, it is clear that sulfonation of peptides is apparently complete (Figure 3.28b). As previously observed for peptides derivatized with SACA, sulfonation with both reagents results in suppression of peptide signal in positive mode MS analysis. Figure 3.28b shows the spectrum of gel band 12 before (upper panel) and after derivatization with SPITC (lower panel). The mass shift of 215 Da, indicative of sulfonation with SPITC is indicated between the spectra. Tryptic fragments with a high molecular weight, > 2200 Da, are generally suppressed upon sulfonation.

The presence of numerous sulfonated peptides in a sample, resulting from different comigrating proteins, precluded the selection of a single peptide for fragmentation (Figure 3.28b, lower spectrum). Lowering the mass window for the selection of a precursor could not eliviate this. Because the fragmentation of multiple peptides in one MS/MS analysis results in

complex spectra with overlapping y-ion series, these fragmentation spectra were not interpreted. Nevertheless, the initial loss of the derivatizing group, a loss of 215 Da for SPITC and 184 Da for SACA, was observed in every MS/MS spectrum acquired during this study, even in those spectra that were to complex for interpretation. This confirms that apparently all peptides were derivatized. In order to avoid the fragmentation of multiple peptides in a single MS/MS experiment, gel bands 16 – 30 were excized narrower and processed as before. The proteins identified in the 30 gel bands are summarized in Table 3.11a & b for SPITC and SACA respectively. The derived peptide sequences are presented in Appendix IV. The longest peptide sequence that was determined was 24 amino acids and on average more then four peptides were matched to each identified protein. On average, two proteins were identified in each spot for SPITC-treated samples. Applying the significance thresholds we used in previous studies, only 3 proteins were reliably identified in SACA-derivatized samples. In all database searches, the highest scores were proteins *Shewanella* species. This confirms the specificity of homology-based database searches using *de novo* determined peptide sequences.

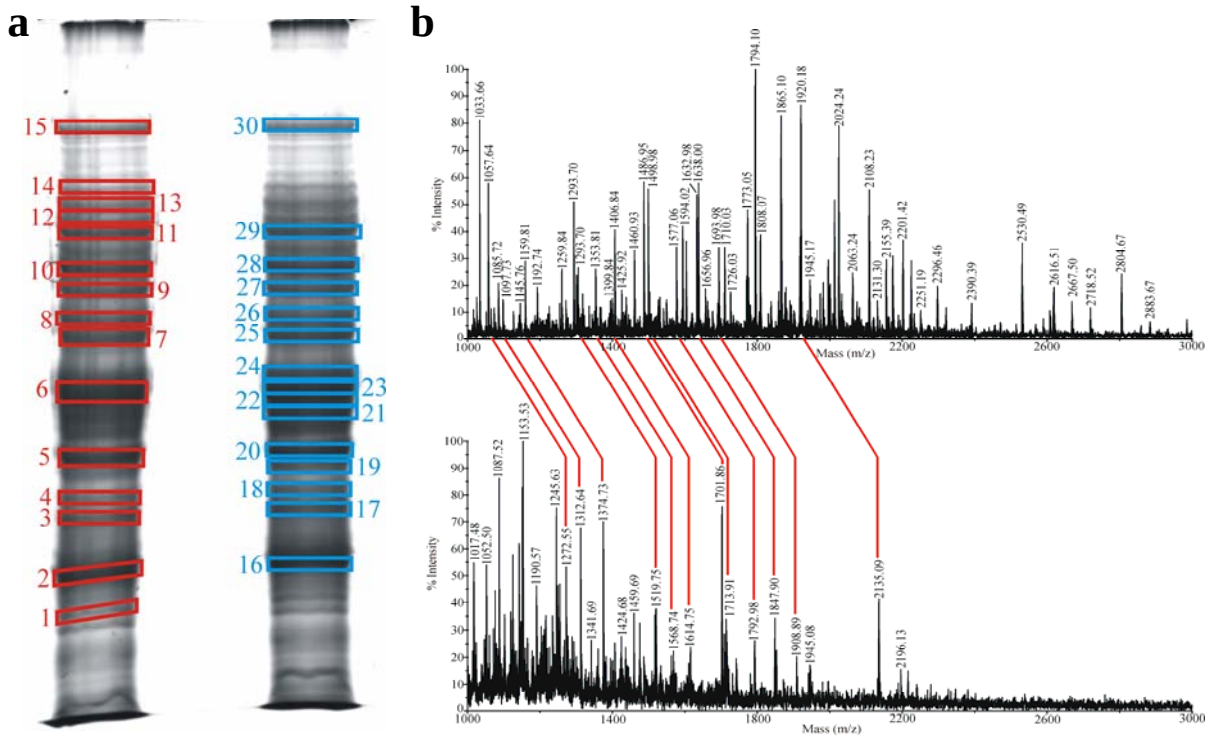


Figure 3.28. a) SDS-PAGE-separated cellular extract of anaerobically grown *Shewanella oneidensis.* The analyzed gel bands are indicated. b) Positive mode MS-spectra of the sample 12 (Figure 3.28a) upper spectrum: peptide mixture after in-gel guanidination of the excized band; lower spectrum: the same sample after sulfonation using SPITC. Mass shifts of 215 Da, indicating sulfonation, are assigned between the two spectra.

Table 3.11a. Results of SPITC derivatization of the spots depicted in Figure 3.28a

Spot[a]	Protein[b]	Identification[c]	Species[c]	FASTS[d]	Scores MS BLAST[e]	MS-Hom.[f]
1	-					
2	24376191	hypothetical protein SO4719	*Shewanella sp.*	7.3e-28	298	270
	24371835	ribosomal protein S3	*Shewanella sp.*	5.9e-14	163	128
	78692243	α-keto acid dehydrog E1 compβ	*Shewanella sp.*	2.8e-10	135	123
3	24375384	outer membrane porin, putative	*Shewanella sp.*	4.5e-66	554	518
	69953123	ketose-bisP aldolase, class-II:	*Shewanella sp.*	1.9e-10	148	132
	24374021	alc dehydrogenase, iron-containing	*Shewanella sp.*	4.0e-07	97	88

4	24375384	outer membrane porin, putative	*Shewanella sp.*	3.8e-45	401	446
	78365797	ketose-bisP aldolase, class-II:	*Shewanella sp.*	1.0e-26	227	239
5	82498382	TEF Tu: Small GTP-bind prot dom	*Shewanella sp.*	7.3e-42	359	334
	78690797	outer membrane porin, putative	*Shewanella sp.*	1.2e-25	243	231
	69953123	ketose-bisphosphate aldolase	*Shewanella sp.*	4.9e-07	93	93
6	24371815	translation elongation factor Tu	*Shewanella sp.*	2e-95	797	742
	77816497	citrate (Si)-synthase	*Shewanella sp.*	1.6e-22	227	208
	24374618	long-chain fatty acid transp prot, put	*Shewanella sp.*	5.9e-05	79	59
7	68545906	IMP dehydrogenase	*Shewanella sp.*	4.0e-18	155	174
	82744041	trigger factor	*Shewanella sp.*	2.3e-18	131	182
	24372021	dihydrolipoamide dehydrogenase	*Shewanella sp.*	4.8e-13	125	154
	82495877	P-enolpyruvate carboxykin	*Shewanella sp.*	2.9e-07	122	92
	82744041	GTPases - transl elong factors	*Shewanella sp.*	8.6e-07	92	109
8	24372295	chaperonin GroEL	*Shewanella sp.*	1.5e-56	496	461
9	24372557	fum reduct flavoprot subu prec	*Shewanella sp.*	1.8e-51	452	413
10	78369113	ribosomal protein S1	*Shewanella sp.*	1.7e-06	112	114
11	78692288	phosphoenolpyruvate synthase	*Shewanella sp.*	5.1E-57	519	429
12	-					
13	-					
14	24374798	P-ribosylformylglycinamidine synth	*Shewanella sp.*	2.0e-20	220	212
15	24374136	hypothetical protein SO2593	*Shewanella sp.*	1.4e-31	348	340
16	24376191	hypothetical protein SO4719	*Shewanella sp.*	9.1e-40	357	324
	78692093	ribosomal protein S3	*Shewanella sp.*	1.1e-25	230	173
17	82744520	ribosomal protein L2	*Shewanella sp.*	3.8e-14	161	186
	82744371	TEF Tu:Small GTP-bind prot dom	*Shewanella sp.*	4.1e-08	119	152
	24373887	α-keto acid dehydrog E1 compβ	*Shewanella sp.*	0.0012	71	84
18	78688396	TEF Tu:Small GTP-bind prot dom	*Shewanella sp.*	3.6e-11	135	144
	69950403	ribose-P-pyrophosphokinase	*Shewanella sp.*	0.0008	82	82
19	68544347	malate dehydrogenase, NAD-dep	*Shewanella sp.*	2.4e-25	252	216
	24371827	translation elongation factor Tu	*Shewanella sp.*	5.1e-10	131	139
	78687894	inorganic diphosphatase	*Shewanella sp.*	0.033	64	65
20	24375384	outer membrane porin, putative	*Shewanella sp.*	1.5e-24	238	242
	24374021	alc dehydroge, iron-containing	*Shewanella sp.*	0.0001	82	88
21	-					
22	-					
23	-					
24	24371827	translation elongation factor Tu	*Shewanella sp.*	8.0e-40	384	356
	77816497	citrate (Si)-synthase	*Shewanella sp.*	1.1e-08	125	136
25	78688058	dihydrolipoamide dehydrogenase	*Shewanella sp.*	4.5e-24	252	256
	63079040	elongation factor-Tu-2	*Shewanella sp.*	2.2e-19	246	193
	78506850	phosphoenolpyruv carboxykin	*Shewanella sp.*	1.3e-09	131	135
	78069867	citrate synthase	*Shewanella sp.*	0.00052	82	93
26	24372295	chaperonin GroEL	*Shewanella sp.*	6.7e-43	415	399
	24373894	glyceraldehyde-3-P-dehydroge	*Shewanella sp.*	3.6e-10	129	132
	24371827	translation elongation factor Tu	*Shewanella sp.*	3.4e-07	113	104
27	24372557	fum reduct flavoprot subu prec	*Shewanella sp.*	2.0e-29	288	301
	24371815	translation elongation factor Tu	*Shewanella sp.*	0.00011	93	91
28	24373579	heat shock protein 90	*Shewanella sp.*	3.9e-23	205	211
	78369113	ribosomal protein S1	*Shewanella sp.*	2.7e-13	152	160
29	78506612	phosphoenolpyruvate synthase	*Shewanella sp.*	3.7e-56	509	655
	78367230	TEF G:Small GTP-bind prot dom	*Shewanella sp.*	1.7e-27	285	282
30	24374136	hypothetical protein SO2593	*Shewanella sp.*	5.8e-65	641	618

Table 3.11b. Results of derivatization of the spots depicted in Figure 3.28a with SACA

Spot[a]	Protein[b]	Identification[c]	Species[c]	Scores FASTS[d]	MS-BLAST[e]	MS-Hom.[f]
3	82497768	outer membrane porin, putative	*Shewanella sp.*	0.033	63	49
4[g]	82497768	outer membrane porin, putative	*Shewanella sp.*	4.8e-13	145	131
6[g]	78688396	TEF Tu:Small GTP-bind prot dom	*Shewanella sp.*	9.3e-15	162	230
7	78506850	P-enolpyruvate carboxykinase	*Shewanella sp.*	0.37	67	60
8	68544820	glyceraldehyde 3-P-dehydroge	*Shewanella sp.*	0.71	78	nr
11	27361244	translation elongation factor G	*Shewanella sp.*	0.0091	69	70
15[g]	78685062	NAD-glutamate dehydrogenase	*Shewanella sp.*	0.00033	74	76
30	24374136	hypothetical protein SO2593	*Shewanella sp.*	0.048	101	89

[a] Spot number according to figure 3.28a
[b] NCBI *Entrez* entries (http://www.ncbi.nih.gov/Entrez/)
[c] Species and identification based on FASTS search results
[d] FASTS score for database searches against the non-redundant NCBI database
[e] MS BLAST score for searches against a nr database at http://dove.embl-Heidelberg.de/Blast2/msblast.html
[f] MS-Homology score against the NCBI on-redundant protein database (Protein Prospector 4.0.5)
[g] Applying the thresholds of previous studies, only these proteins were identified after SACA-derivatization

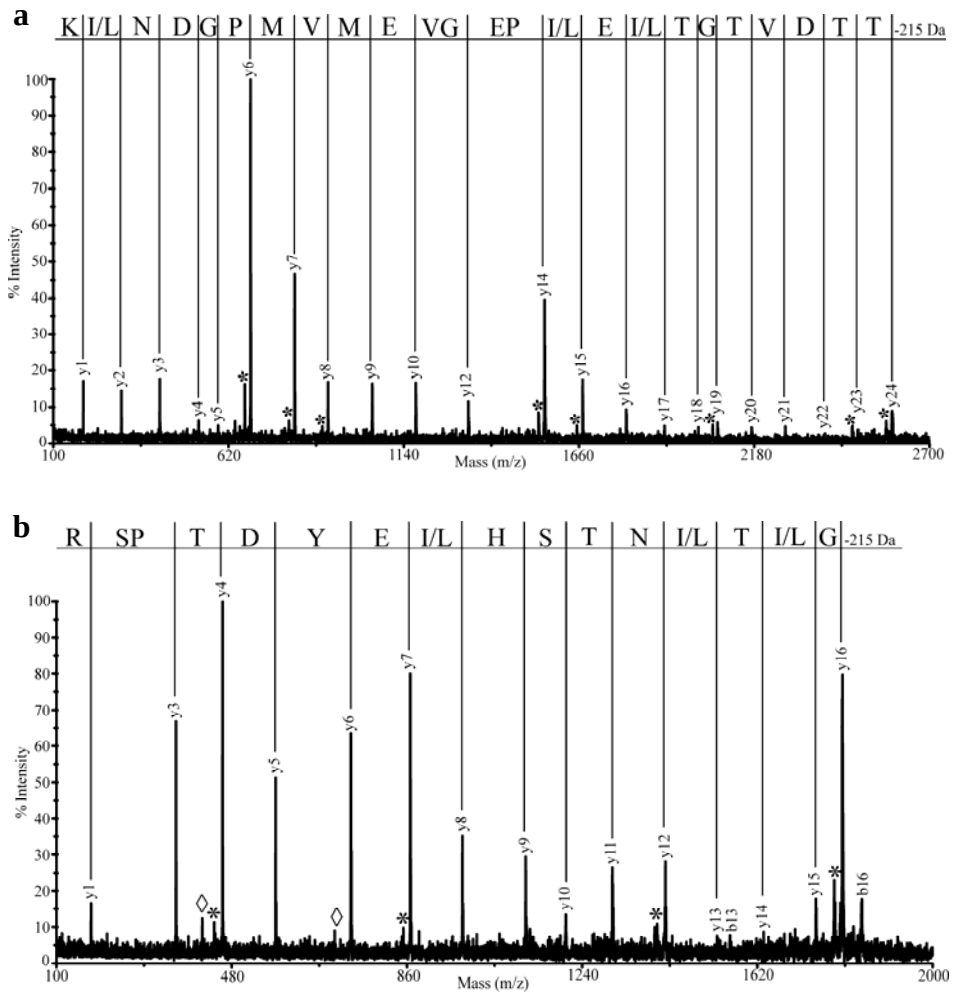


Figure 3.29. Fragmentation spectra of guanidinated and SPITC-derivatized peptides from gel band 6 (Figure 3.28a). Fragmentation spectrum of the peptide precursor at 2802.28 Da (a) and of the precursor at 2018,81 Da (b). The loss of the derivative is observed by the 215 Da mass shift and y-ions are indicated. (y-17)-ions resulting from the neutral loss of NH_3 are labeled as * and internal ions with a diamond. The determined peptide sequence is indicated in the one-letter code.

Fragmentation spectra of SPITC-derivatized peptides are similar to those of peptides derivatized with SACA. The ladder-like series of y-ions allow easy manual sequence determination; nevertheless, we observed more small gaps of two amino acids in these fragmentation spectra (Figure 3.29). All the observed gaps in this study have the sequence Gly-Xxx or Pro-Xxx. The presence of gaps at these residues is in agreement with current knowledge on preferential fragmentation pathways. Therefore, in most instances the sequence of these gaps could be reliably determined (Figure 3.29). Mass differences of 113 Da between consecutive y-ions, indicating isoleucine or leucine, were submitted as 'L' in database searches. Therefore, the ambiguous identification of Ile and Leu only had to be considered during MS-Homology-searches, because an accepted number of non-identical amino acids in each peptide must be specified. Since no unmodified lysine residues were observed, the isobaric amino acids lysine and glutamine were distinguished unambiguously.

Discussion

The aim of the current study was to compare the sulfonation of peptides with SACA or SPITC. Therefore, exactly the same sample was used for both approaches. For derivatization with 2-sulfobenzoic acid cyclic anhydride the optimized protocol was used (Part 3.3.1 & 3.3.2). SPITC-derivatization was performed according to a protocol described by Wang *et al* with minor modification (Wang *et al*, 2004). In order to avoid increased sodium contamination of the sample, a volatile 20 mM $NH4HCO_3$-buffer has been used instead of the 20 mM $NaHCO_3$-buffer used by Wang *et al*.

Compared to our previous study (Sergeant *et al*, 2005) a ten times lower initial sample amount was used (50 µg compared to 500 µg). Because the peptide extract was divided between the two experiments, an even lower sample amount was used for each derivatization experiment. The use of THF as cosolvent and its effect on the sensitivity of MALDI MS analysis is not reported in literature nor are exhaustive studies on the effects of ACN. However, a study published in 1996 indicated that the sensitivity of MALDI MS analysis is influenced by the composition of the solvent wherein sample is applied on the target plate (Cohen *et al*, 1996). One possible explanation is that peptides are less soluble in solutions that contain high concentrations of organic solvents; e.g. 25% THF and 37.5% ACN for SACA-derivatized peptides compared to 25% ACN for SPITC-derivatized peptides.

After SPITC-derivatization, several gel bands resulted in the identification of multiple proteins with a maximum of 5 proteins identified in band 7 (Table 3.11a). Nevertheless, *de novo* sequence determination using the approach described here is limited to relatively simple mixtures. The difficulties encountered in selecting single peptides for MS/MS from sample 12 (Figure 3.28b) illustrate that the number of proteins that can be identified in a sample is limited. Although apparently complete sulfonation was observed, the difficulties encountered in the interpretation of MS/MS spectra from multiple peptides abolish the advantage of peptide sulfonation. Apart from band 12, no single peptide could be selected for 5 of the other samples; no efforts were done to interpreter the complex spectra that resulted from the fragmentation of multiple peptides. LC separation of the sulfonated peptides prior to MALDI analysis will probably result in the identification of proteins from these samples (Lee *et al*, 2004b; Flensburg *et al*, 2005).

SPITC-derivatized peptides readily fragment and y-ions are the most abundant ions in the fragmentation spectra (Figure 3.29). However, at this point it cannot be concluded if the

same rules, described for preferential fragmentation of SACA-derivatized peptides (Samyn *et al*, 2004), can be applied to SPITC-derivatized peptides. In a Master of Science thesis by Lene Jensen (http://www.sdu.dk/Nat/bmb/Newsletter/BMB,news1/lene.htm) sulfonation with CAF and SPITC were compared. It was concluded that the use of both reagents resulted in spectra that are easy to interprete, allowing straightforward *de novo* sequence determination. However, it was also concluded that no specific fragmentation rule can be applied to the fragmentation spectra of SPITC-derivatized peptides. In the study presented here, some preferential fragmentation patterns were observed, for instance low or absent peaks resulting from the fragmentation of peptide bonds C-terminal to Gly or Pro (Figure 3.29), although they cannot be as strictly superimposed on fragmentation spectra as for SACA derivatized peptides.

The results presented here clearly indicate that the use of SPITC is a superior alternative to the use of SACA for N-terminal sulfonation of peptides. Furthermore, the application of the sulfonation reagent in non-toxic aqueous environment makes the in-gel guanidination protocol, followed by sulfonation, easier to perform and more amendable to automation.

3.3.3.2. Automation of the in-gel guanidination protocol: initial results

Automation of steps in protocols expedites proteome analysis and as such is a prerequisite for high-throughput analyses (Lopez, 2000). In the in-gel guanidination protocol most steps, except the running of 2D-PAGE gels, can be automated. Here, we describe initial results in which the wet-lab steps of our protocol, starting from spot picking to sample preparation for MS analysis, are automated. In general, automation of individual steps increases the overall accuracy, reproducibility and throughput. Furthermore, reducing human interference to a minimum eliminates the possibility of human errors and common contaminants (e.g. keratin) (Houthaeve *et al*, 1997).

This work was performed in collaboration with Dr. Jenny Renaut and Dr. Jean-François Hausman from the Centre de Reserche Public Gabriel Lippman (CRPGL) Luxembourg. The instrument is an Ettan™ Spot Handling Workstation (GE Healthcare) in which all sample processing steps are performed on a computer-controlled platform (Figure 3.30). The Ettan workstation is a dedicated robotic system that integrates all the steps needed in standard trypsin digestion protocols, while some flexibility is allowed in each step. The highly efficient spot picker can be equipped with a 1.4 or a 2 mm picker head. The incubator can be set at any temperature from 20 to 65°C, allowing in-gel guanidination to be automatically performed. The Ettan workstation is computer-controlled and the interface with the DeCyder™ image analysis software ensures easy transfer of gel scans and 'pick'-lists.

Figure 3.30. The Ettan™ Spot Handling Workstation from GE Healthcare (CRPGL, Luxembourg). Spot picking, destaining, trypsin digestion, peptide extraction and sample spotting for MALDI MS analysis are integrated. The blow-ups show the spot picker and the robotic arm that transports the samples between the different modules.

Despite the flexibility that can be afforded in this system, we noticed that the entire protocol cannot be performed in a single cycle. The software does not allow to perform an incubation step after drying of the extracted tryptic peptides. Therefore, sulfonation reagent was manually added in a first study (results not shown). After incubation, the samples could be spotted automatically. Although the protocol was not fully automated, the results of this initial study demonstrated that automation of the first part of the protocol, including the in-gel

guanidination, is possible. To avoid the manual step, the protocol was divided in two cycles that are automatically run one after the other (Figure 3.31).

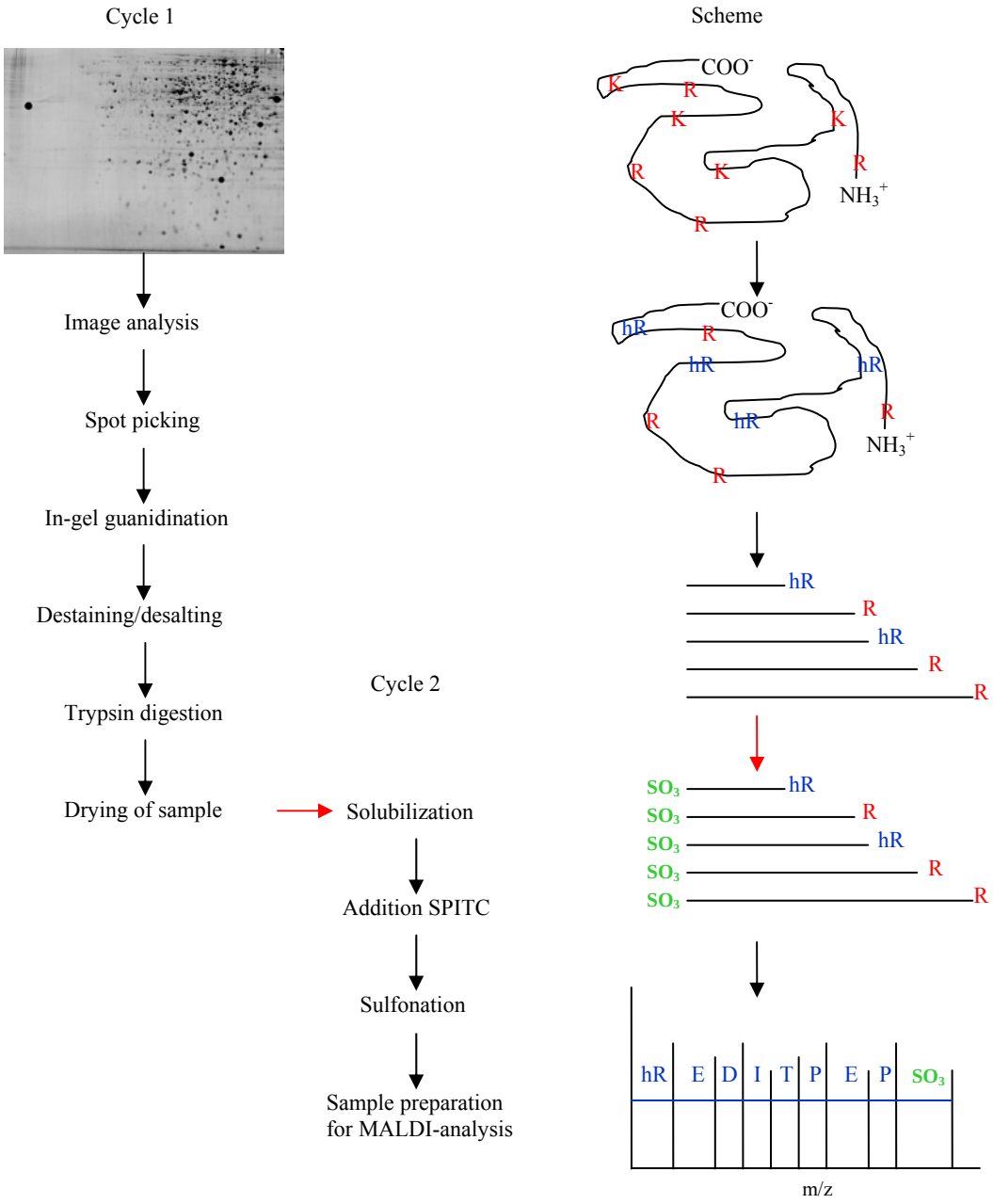


Figure 3.31. The final protocol as programmed for automated in-gel guanidination, trypsin digestion and sulfonation. The protocol is split in two cycles that are run consecutively; performing these two cycles results in fully automated MALDI sample preparation starting from 2D-PAGE separated proteins. hR: homoarginine; SO_3^- : sulfophenylthiocarbamoyl derivatized N-terminal amino group.

As depicted in Figure 3.31, the first cycle includes all steps up to drying of the extracted tryptic peptides, including in-gel guanidination. An adjusted cycle is used for sulfonation. 6 µl of a SPITC-solution (12.5 mM ammonium bicarbonate in MQ containing 10 mg SPITC/ml) is added to the dried sample and the microtiter plate is incubated at 37°C. After an incubation of 30 minutes, 0.1 µl derivatization buffer is added, to mimic resolubilization of dried peptide extract, and the samples are automatically spotted.

Proteins isolated from *Musa balbisiani* (Laboratory for Tropical Crop Improvement, KUL) were separated by 2D-PAGE and used to test the applicability of the automated protocol. Initially, 35 spots were selected (Figure 3.32) and submitted to the protocol. The major difference in running 2D-PAGE gels for use with the automated platform is that reference markers had to be inserted in the gel as positional references for the spot picker. To avoid shrinking of the gel and distortion of the gel image during scanning and spot picking, the gels are glued to one of the glass plates. Spots with a diameter of 2 mm were picked and processed according to the protocol depicted in Figure 3.31.

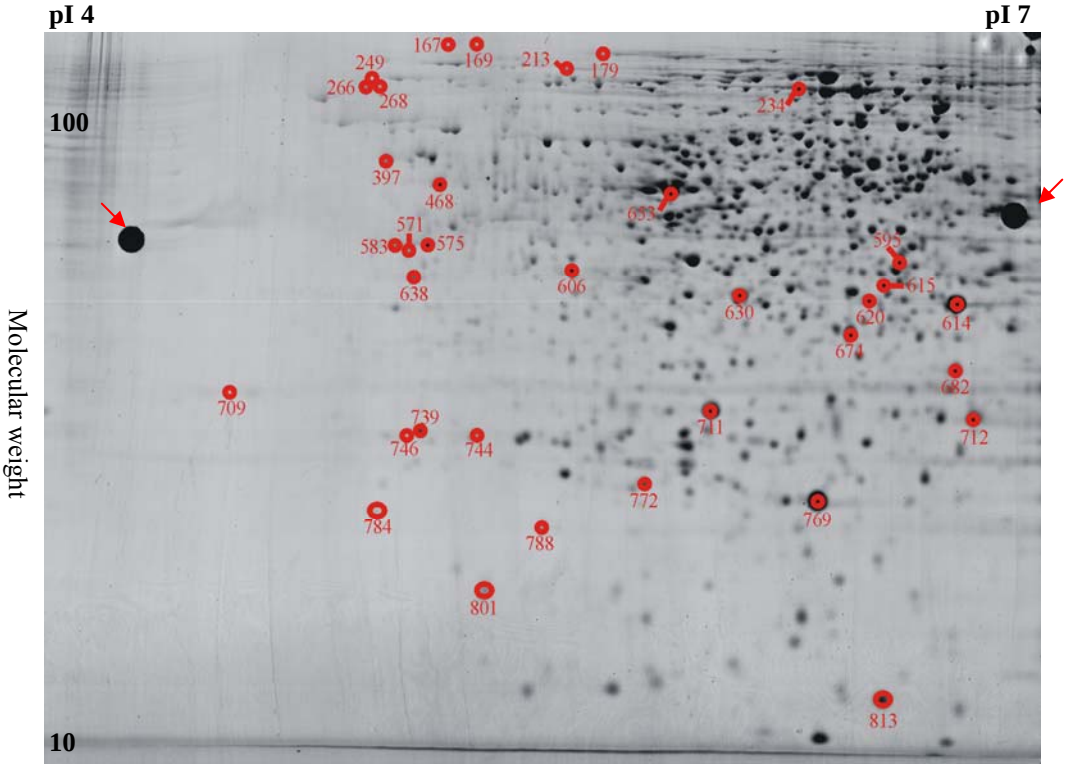


Figure 3.32. Gel image of the *Musa balbisiani* gel used to test the automated protocol represented in Figure 3.31, the spots that were picked are indicated. The arrows point at the reference markers used for positional calibration of the spot picker.

Because sulfonation is expected to be nearly complete (Part 3.3.3.1), the analysis of guanidinated sample prior to sulfonation was omitted. Sulfonated peptides were detected in negative ion mode (Samyn *et al*, 2004; Sergeant *et al*, 2005). As decribed before, selection of the corresponding protonated precursor ion for TOF/TOF analysis (positive mode) resulted in the formation of complete series of y fragment ions (Figure 3.33). A similar approach was recently used in a study of cysteic acid containing peptides (Dai *et al*, 2005). The selective detection of sulfonated peptides during negative mode MS analysis is illustrated in Figure 3.33a & b. Figure 3.33a is the positive mode MS spectrum of the sulfonated peptides from spot 571; the same peptides, although at a lower m/z-value ($\Delta m/z$ = -2Da) are observed in the negative mode MS spectrum (Figure 3.33b). Some sulfonated peptides, not observed in the positive mode spectrum, are observed as deprotonated peptides in the negative mode. The positive mode fragmentation spectra of two of these peptides, precursors selected at 2069.11 and 2225.30 Da corresponding to the peaks at 2067.11 and 2223.30 in negative mode, are shown in Figure 3.33c and 3.33d respectively.

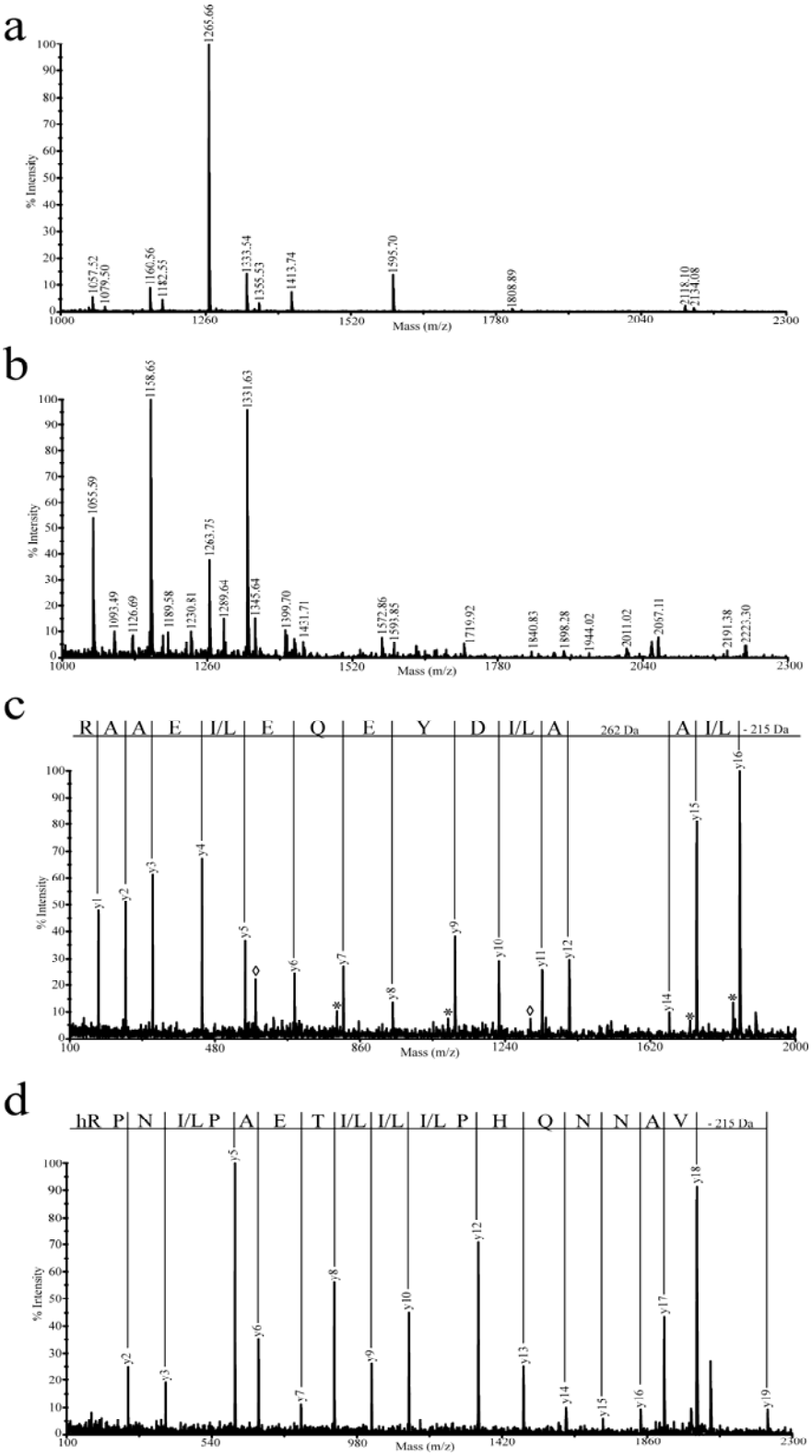


Figure 3.33. Mass spectra of the peptides after automated processing of spot 571 (Figure 3.32) according to the protocol depicted in Figure 3.31. a) Positive mode MS analysis. b) negative mode MS analysis of the same sample, c) fragmentation spectra of the precursor at 2069.11 Da and d) of the precursor at 2225.30 Da observed in b) at respectively 2067.11 and 2223.30 Da. Loss of the sulfonating group is indicated as -215 Da. (y-17)-ions resulting from the neutral loss of NH_3 are labeled with *, internal ions with diamonds.

We were able to identify proteins in 28 of the 35 spots (Table 3.12, Appendix V). Some proteins were identified in multiple spots, e.g. actin in spots 571 & 575 and

triosephosphate isomerase like-protein in spots 744 & 746. In every database search, the proteins having the highest homology to the query are database entries from the evolutionary clade of the green plants, Viridiplantae. One exception is the protein 3-isopropylmalate dehydrogenase, identified in spot 583, that is involved in leucine biosynthesis in the proteobacterium *Pemphigus spyrothecae* (query; LEAAVLNTLNR and ELTGGLYFGKPR). The species indicated in the table corresponds to the hit with the highest score using FASTS. When the same sequences were submitted in a MS-Homology search, the highest hit was a homologous protein from rice, with matching sequences **V**EAAV**TE**TLN**N** and ELTGG**I**YFG**Q**PR. Furthermore, a conventional BLAST search using only the sequence ELTGGLYFGKPR revealed that this peptide sequence is highly conserved among the 3-isopropylmalate dehydrogenase proteins from bacteria and plants.

Table 3.12. Proteins identified in the spots depicted in Figure 15 using the automated approach

Spot[a]	Protein[b]	Identification FASTS[c]	Species[c]	FASTS[d]	MS BLAST[e]	MS-Hom[f]
167	-					
169	-					
179	-					
213	-					
234	11066033	cytosolic aconitase	*Nicotiana tabacum*	1.8e-07	71	101
249	-					
266	77556324	putative heat-shock protein	*Oryza sativa*	1.7e-18	228	198
268	-					
397	24637539	heat shock protein 60	*Prunus dulcis*	8.1e-17	140	177
453	37531422	putative enolase	*Oryza sativa*	1.4e-32	293	288
	2645893	F1 ATPase a-subunit	*Panax ginseng*	3.2e-09	156	139
468	22273	enolase	*Zea mays*	2.2e-39	381	434
571	33339126	actin	*Musa acuminate*	4.1e-40	395	365
575	1498395	actin	*Zea mays*	6.6e-36	320	351
583	20513166	3-isopropylmalate dehydroge	*Pemphigus spyrothecae*	9.7e-08	103	114
	3738259	cytos phosphoglycerate kin. 1	*Populus nigra*	1.6e-06	107	108
595	50251688	putative aspartate transaminase	*Oryza sativa*	9.5e-12	144	140
606	60101357	glutamine synthetase	*Vigna radiate*	3.6e-14	157	134
614	50940085	putative r40c1 protein	*Oryza sativa*	3.2e-11	169	188
615	1620972	L-lactate dehydrogenase	*Lycopersicon esculentum*	1.1e-14	187	169
620	1658313	osr40g2	*Oryza sativa*	5.8e-10	133	129
630	10798652	malate dehydrogenase	*Nicotiana tabacum*	1.6e-26	259	261
638	30060226	1-aminocycloprop oxid	*Elaeis guineensis*	1.1e-15	197	238
674	76559896	TPA: isoflav reduc-like prot 6	*Vitis vinifera*	2.9e-30	308	320
682	15705988	endochitinase	*Musa acuminate*	2.6e-23	226	213
709	7739434	14-3-3-like protein	*Capsicum annuum*	4.2e-22	221	179
711	41818408	class III acidic chitinase	*Musa acuminate*	1.2e-20	237	215
712	1668706	atran2	*Arabidopsis thaliana*	9.1e-26	269	241
739	2586151	ripening-associated protein	*Musa acuminate*	8.6e-17	164	179
744	76573375	triosephosphate isom-like prot	*Solanum tuberosum*	6.9e-15	163	166
746	76573375	triosephosphate isom-like prot	*Solanum tuberosum*	7.9e-17	174	163
769	47575681	abscisic stress ripening prot	*Musa acuminate*	8.6e-06	62	96
772	601871	manganese-superoxide dismut	*Oryza sativa*	8.2e-07	102	121

784	3492854	mitoch small heat shock prot	*Lycopersicon esculentum*	3.0e-06	66	na
788	34334012	cytos glutathione peroxidase	*Triticum monococcum*	9.1e-24	225	223
801	-					
813	47026989	nucleoside diphosphate kinase	*Hyacinthus orientalis*	8.4e-25	234	254

[a] spot number according to Figure 3.32
[b] NCBI *Entrez* entries (http://www.ncbi.nih.gov/Entrez/)
[c] Species and identification based on FASTS search results
[d] FASTS score for database searches against the non-redundant NCBI database
[e] MS BLAST score for searches against a nr database at http://dove.embl-Heidelberg.de/Blast2/msblast.html
[f] MS-Homology score against the NCBI on-redundant protein database (Protein Prospector 4.0.5

The results reported here demonstrate that automation of the in-gel guanidination protocol is a viable method for cross-species identification using *de novo* determined sequences. However, we realize that automation of this protocol is only a first step towards a high throughput method for cross-species protein identification. To attain this goal, automation of the spectral interpretation and database searches will be required.

References

Aebersold, R.; Mann, M. (**2003**). Mass spectrometry-based proteomics. *Nature* **422**(6928): 198-207.

Agrawal, G.K.; Rakwal, R. (**2006**). Rice proteomics: a cornerstone for cereal food crop proteomes. *Mass Spectrom Rev* **25**(1): 1-53.

Ali, G.M.; Komatsu, S. (**2006**). Proteomic analysis of rice leaf sheath during drought stress. *J. Proteom. Res.*: DOI 10.1021/pr050291q.

Altschul, S.F.; Boguski, M.S.; Gish, W.*; et al.* (**1994**). Issues in searching molecular sequence databases. *Nat Genet* **6**(2): 119-29.

Amiour, N.; Merlino, M.; Leroy, P.*; et al.* (**2002**). Proteomic analysis of amphiphilic proteins of hexaploid wheat kernels. *Proteomics* **2**(6): 632-41.

Arnott, D.; Henzel, W.J.; Stults, J.T. (**1998**). Rapid identification of comigrating gel-isolated proteins by ion trap-mass spectrometry. *Electrophoresis* **19**(6): 968-80.

Bailey, T.H.; Laskin, J.; Futrell, J.H. (**2003**). Energetics of selective cleavage at acidic residues studied by time- and energy-resolved surface-induced dissociation in FT-ICR MS. *Int J Mass Spectrom* **222**(1-3): 313-27.

Ballard, K.D.; Gaskell, S.J. (**1993**). Dehydration of peptide [M + H]+ ions in the gas phase. *J Am Soc Mass Spectrom* **4**(6): 477-81.

Bartlet-Jones, M.; Jeffery, W.A.; Hansen, H.F.*; et al.* (**1994**). Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Commun Mass Spectrom* **8**(9): 737-42.

Bauer, M.D.; Sun, Y.; Keough, T.*; et al.* (**2000**). Sequencing of sulfonic acid derivatized peptides by electrospray mass spectrometry. *Rapid Commun Mass Spectrom* **14**(10): 924-9.

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 Commun Mass Spectrom* **14**(23): 2147-53.

Beardsley, R.L.; Reilly, J.P. (**2002**). Optimization of guanidination procedures for MALDI mass mapping. *Anal Chem* **74**(8): 1884-90.

Beardsley, R.L.; Sharon, L.A.; Reilly, J.P. (**2005**). Peptide de novo sequencing facilitated by a dual-labeling strategy. *Anal Chem* **77**(19): 6300-9.

Bergman, T. **(2000)**. Ladder sequencing. ed. Jolles, P. Jornvall, H. *Proteomics in functional genomics.* Basel/Switzerland, Birkhaüser Verlag. **88:** 133-44.

Beye, M.; Poch, A.; Burgtorf, C.*; et al.* (**1998**). A gridded genomic library of the honeybee (Apis mellifera): a reference library system for basic and comparative genetic studies of a hymenopteran genome. *Genomics* **49**(2): 317-20.

Bocharov, E.V.; Sobol, A.G.; Pavlov, K.V.*; et al.* (**2004**). From structure and dynamics of protein L7/L12 to molecular switching in ribosome. *J Biol Chem* **279**(17): 17697-706.

Bonetto, V.; Bergman, A.C.; Jornvall, H.*; et al.* (**1997**). C-terminal sequence analysis of peptides and proteins using carboxypeptidases and mass spectrometry after derivatization of Lys and Cys residues. *Anal Chem* **69**(7): 1315-9.

Boudart, G.; Jamet, E.; Rossignol, M.*; et al.* (**2005**). Cell wall proteins in apoplastic fluids of Arabidopsis thaliana rosettes: identification by mass spectrometry and bioinformatics. *Proteomics* **5**(1): 212-21.

Brancia, F.L.; Oliver, S.G.; Gaskell, S.J. (**2000**). Improved matrix-assisted laser desorption/ionization mass spectrometric analysis of tryptic hydrolysates of proteins following guanidination of lysine-containing peptides. *Rapid Commun Mass Spectrom* **14**(21): 2070-3.

Brancia, F.L.; Butt, A.; Beynon, R.J.*; et al.* (**2001**). A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis* **22**(3): 552-9.

Brancia, F.L.; Montgomery, H.; Tanaka, K.*; et al.* (**2004**). Guanidino labeling derivatization strategy for global characterization of peptide mixtures by liquid chromatography matrix-assisted laser desorption/ionization mass spectrometry. *Anal Chem* **76**(10): 2748-55.

Breci, L.A.; Tabb, D.L.; Yates, J.R., 3rd*; et al.* (**2003**). Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal Chem* **75**(9): 1963-71.

Bunk, D.M.; MacFarlane, R.D. (**1993**). Derivatization to enhance sequence-specific fragmentation of peptides and proteins. *Int J Mass spectr ion process* **126**: 123-36.

Burlet, O.; Yang, C.; Gaskell, S.J. (**1992**). Influence of cysteine to cysteic acid oxidation on the collision-activated decomposition of protonated peptides: evidence for intraionic interactions. *J Am Soc Mass Spectrom* **3**(4): 337-44.

Burlet, O.; Yang, C.; J.R., G.*; et al.* (**1995**). Tandem mass spectrometric characterization of a specific cysteic acid residue in oxidized human apoprotein B-100. *J Am Soc Mass Spectrom* **6**(4): 242-47.

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

Campbell, J.M. (**2003**). Mapping the properties of center of mass collision energy on a MALDI TOF/TOF mass spectrometer - fundamentals and applications. *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics, Montreal*.

Cantin, G.T.; Yates, J.R., 3rd (**2004**). Strategies for shotgun identification of post-translational modifications by mass spectrometry. *J Chromatogr A* **1053**(1-2): 7-14.

Carpentier, S.C.; Witters, E.; Laukens, K.*; et al.* (**2005**). Preparation of protein extracts from recalcitrant plant tissues: an evaluation of different methods for two-dimensional gel electrophoresis analysis. *Proteomics* **5**(10): 2497-507.

Chait, B.T.; Wang, R.; Beavis, R.C.*; et al.* (**1993**). Protein ladder sequencing. *Science* **262**(5130): 89-92.

Chang, W.W.; Huang, L.; Shen, M.*; et al.* (**2000**). Patterns of protein synthesis and tolerance of anoxia in root tips of maize seedlings acclimated to a low-oxygen environment, and identification of proteins by mass spectrometry. *Plant Physiol* **122**(2): 295-318.

Chen, X.; Turecek, F. (**2005**). Simple B ions have cyclic oxazolone structures. A neutralization-reionization mass spectrometric and computational study of oxazolone radicals. *J Am Soc Mass Spectrom* **16**(12): 1941-56.

Chong, P.K.; Wright, P.C. (**2005**). Identification and characterization of the Sulfolobus solfataricus P2 proteome. *J Proteome Res* **4**(5): 1789-98.

Clauser, K.R.; Baker, P.; Burlingame, A.L. (**1999**). Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* **71**(14): 2871-82.

Cohen, S.L.; Chait, B.T. (**1996**). Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins. *Anal Chem* **68**(1): 31-7.

Cordwell, S.J.; Wilkins, M.R.; Cerpa-Poljak, A.*; et al.* (**1995**). Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionisation/time-of-flight mass spectrometry and amino acid composition. *Electrophoresis* **16**(3): 438-43.

Cotter, R.J.; Ramirez, S.M.; Soloski, M.J. (**2001**). Guanidination of Class I peptides containing lysine to resolve a sequence ambiguity. *Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics, Chicago, Illinois*.

Cox, K.A.; Gaskell, S.J.; Morris, M.*; et al.* (**1996**). Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions. *J Am Soc Mass Spectrom* **7**(6): 522-31.

Cramer, R.; Corless, S. (**2001**). The nature of collision-induced dissociation processes of doubly protonated peptides: comparative study for the future use of matrix-assisted laser desorption/ionization on a hybrid quadrupole time-of-flight mass spectrometer in proteomics. *Rapid Commun Mass Spectrom* **15**(22): 2058-66.

Cupo, P.; El-Deiry, W.; Whitney, P.L.*; et al.* (**1980**). Stabilization of proteins by guanidination. *J Biol Chem* **255**(22): 10828-33.

Czeszak, X.; Morelle, W.; Ricart, G.*; et al.* (**2004**). Localization of the O-glycosylated sites in peptides by fixed-charge derivatization with a phosphonium group. *Anal Chem* **76**(15): 4320-4.

Dai, J.; Wang, J.; Zhang, Y.*; et al.* (**2005**). Enrichment and Identification of Cysteine-Containing Peptides from Tryptic Digests of Performic Oxidized Proteins by Strong Cation Exchange LC and MALDI-TOF/TOF MS *Anal Chem* **77**(23): 7594-604.

Denby, K.; Gehring, C. (**2005**). Engineering drought and salinity tolerance in plants: lessons from genome-wide expression profiling in Arabidopsis. *Trends Biotechnol* **23**(11): 547-52.

Dongré, A.R.; Jones, J.L.; Somogyi, A.*; et al.* (**1996**). Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: evidence for the mobile proton model *J Am Chem Soc* **118**(35): 8365 -74.

Edman, P.; Begg, G. (**1967**). A protein sequenator. *Eur J Biochem* **1**(1): 80-91.

Eng, J.K.; McCormack, A.L.; Yates, J.R.I. (**1994**). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**(11): 976-89.

Fischer, B.; Roth, V.; Roos, F.*; et al.* (**2005**). NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal Chem* **77**(22): 7265-73.

Fleischmann, R.D.; Adams, M.D.; White, O.*; et al.* (**1995**). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**(5223): 496-512.

Flensburg, J.; Tangen, A.; Prieto, M.*; et al.* (**2005**). Chemically-assisted fragmentation combined with multi-dimensional liquid chromatography and matrix-assisted laser desorption/ionization post source decay, matrix-assisted laser desorption/ionization tandem time-of flight or matrix-assisted laser desorption/ionization tandem mass spectrometry for improved sequencing of tryptic peptides. *Eur J Mass Spectrom* **11**(2): 169-79.

Frank, A.; Pevzner, P. (**2005**). PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **77**(4): 964-73.

Frison, E.; Sharrock, S. (**1998**). The economical, social and nutritional importance of banana in the world. In Bananas and Food Security. . *Proceedings of the International Symposium on Bananas and Food Security, Douala, Cameroun*: 21-35.

Gaskell, S.J.; Reilly, M.H.; Porter, C.J. (**1988**). Collisionally activated decomposition of leucine-enkephalin and analogues using a hybrid tandem mass spectrometer. *Rapid Commun Mass Spectrom* **2**(7): 142-5.

Gevaert, K.; De Mol, H.; Verschelde, J.L.*; et al.* (**1997**). Novel techniques for identification and characterization of proteins loaded on gels in femtomole amounts. *J Protein Chem* **16**(5): 335-42.

Gevaert, K.; Demol, H.; Martens, L.*; et al.* (**2001**). Protein identification based on matrix assisted laser desorption/ionization-post source decay-mass spectrometry. *Electrophoresis* **22**(9): 1645-51.

Glasauer, S.; Langley, S.; Beveridge, T.J. (**2002**). Intracellular iron minerals in a dissimilatory iron-reducing bacterium. *Science* **295**(5552): 117-9.

Grossmann, J.; Roos, F.F.; Cieliebak, M.*; et al.* (**2005**). AUDENS: a tool for automated peptide de novo sequencing. *J Proteome Res* **4**(5): 1768-74.

Gu, C.; Tsaprailis, G.; Breci, L.*; et al.* (**2000**). Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal Chem* **72**(23): 5804-13.

Gu, C.G.; Somogyi, A.; Wysocki, V.H.*; et al.* (**1999**). Fragmentation of protonated oligopeptides XLDVLQ (X=L, H, K or R) by surface induced dissociation: additional evidence for the 'mobile proton' model. *Anal Chim Acta* **397**(1-3): 247-56.

Gu, S.; Pan, S.; Bradbury, E.M.*; et al.* (**2002**). Use of deuterium-labeled lysine for efficient protein identification and peptide de novo sequencing. *Anal Chem* **74**(22): 5774-85.

Gu, S.; Pan, S.; Bradbury, E.M.*; et al.* (**2003**). Precise peptide sequencing and protein quantification in the human proteome through in vivo lysine-specific mass tagging. *J Am Soc Mass Spectrom* **14**(1): 1-7.

Gygi, S.P.; Rist, B.; Gerber, S.A.*; et al.* (**1999**). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**(10): 994-9.

Habermann, B.; Oegema, J.; Sunyaev, S.*; et al.* (**2004**). The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol Cell Proteomics* **3**(3): 238-49.

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

Halligan, B.D.; Ruotti, V.; Twigger, S.N.*; et al.* (**2005**). DeNovoID: a web-based tool for identifying peptides from sequence and mass tags deduced from de novo peptide sequencing by mass spectroscopy. *Nucleic Acids Res* **33**(Web Server issue): W376-81.

Hara, H.; Nishi, T.; Kasai, T. (**1995**). A protein less sensitive to trypsin, guanidinated casein, is a potent stimulator of exocrine pancreas in rats. *Proc Soc Exp Biol Med* **210**(3): 278-84.

Harrison, A.G.; Yalcin, T. (**1997**). Proton mobility in protonated amino acids and peptides. *Int J Mass Spectr Ion Process* **165**: 339-47.

Harrison, A.G.; Young, A.B. (**2004**). Fragmentation of protonated oligoalanines: amide bond cleavage and beyond. *J Am Soc Mass Spectrom* **15**(12): 1810-9.

Hellman, U.; Bhikhabhai, R. (**2002**). Easy amino acid sequencing of sulfonated peptides using post-source decay on a matrix-assisted laser desorption/ionization time-of-flight mass spectrometer equipped with a variable voltage reflector. *Rapid Commun Mass Spectrom* **16**(19): 1851-9.

Henikoff, S.; Henikoff, J.G. (**1992**). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**(22): 10915-9.

Heredia-Langner, A.; Cannon, W.R.; Jarman, K.D.*; et al.* (**2004**). Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinformatics* **20**(14): 2296-304.

Hjerno, K.; Matthiesen, R.; Roepstorff, P. (**2005**). Improving the confidence of protein identification by combining MALDI-MS and MS/MS data *HUPO 4th Annual World Congress, Munich. Mol. Cell. Proteomics* **4**(8): S302.

Horn, D.M.; Zubarev, R.A.; McLafferty, F.W. (**2000**). Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc Natl Acad Sci U S A* **97**(19): 10313-7.

Houthaeve, T.; Gausepohl, H.; Ashman, K.*; et al.* (**1997**). Automated protein preparation techniques using a digest robot. *J Protein Chem* **16**(5): 343-8.

Huang, Y.; Wysocki, V.H.; Tabb, D.L.*; et al.* (**2002**). The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Int J Mass Spectrom* **219**: 233-44.

Huang, Z.H.; Wu, J.; Roth, K.D.*; et al.* (**1997**). A picomole-scale method for charge derivatization of peptides for sequence analysis by mass spectrometry. *Anal Chem* **69**(2): 137-44.

Hunt, D.F.; Yates, J.R., 3rd; Shabanowitz, J.*; et al.* (**1986**). Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci U S A* **83**(17): 6233-7.

Johnson, R.S.; Martin, S.A.; Biemann, K. (**1988**). Collision-Induced Fragmentation of (M+H)+ Ions of Peptides. Side Chain Specific Sequence Ions. *Int. J. Mass Spectrom. Ion Processes* **86**: 137-54.

Johnson, R.S.; Krylov, D.; Walsh, K.A. (**1995**). Proton mobility within electrosprayed ions. *J Mass Spectrom* **30**: 386-7.

Jones, J.L.; Dongre, A.R.; Somogyi, A.*; et al.* (**1994**). Sequence dependence of peptide fragmentation efficiency curves determined by electrospray ionization/surface-induced dissociation mass spectrometry. *J Am Chem Soc* **116**(18): 8368-9.

Jorge, I.; Navarro, R.M.; Lenz, C.*; et al.* (**2005**). The holm oak leaf proteome: analytical and biological variability in the protein expression level assessed by 2-DE and protein identification tandem mass spectrometry de novo sequencing and sequence similarity searching. *Proteomics* **5**(1): 222-34.

Jorgensen, T.J.; Bache, N.; Roepstorff, P.*; et al.* (**2005a**). Collisional activation by MALDI tandem time-of-flight mass spectrometry induces intramolecular migration of amide hydrogens in protonated peptides. *Mol Cell Proteomics* **4**(12): 1910-9.

Jorgensen, T.J.; Gardsvoll, H.; Ploug, M.*; et al.* (**2005b**). Intramolecular migration of amide hydrogens in protonated peptides upon collisional activation. *J Am Chem Soc* **127**(8): 2785-93.

Kapp, E.A.; Schutz, F.; Reid, G.E.*; et al.* (**2003**). Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem* **75**(22): 6251-64.

Karty, J.A.; Ireland, M.M.; Brun, Y.V.*; et al.* (**2002**). Defining absolute confidence limits in the identification of Caulobacter proteins by peptide mass mapping. *J Proteome Res* **1**(4): 325-35.

Kaufmann, R.; Chaurand, P.; Kirsch, D.*; et al.* (**1996**). Post-source decay and delayed extraction in matrix-assisted laser desorption/ionization-reflectron time-of-flight mass spectrometry. Are there trade-offs? *Rapid Commun Mass Spectrom* **10**(10): 1199-208.

Keough, T.; Youngquist, R.S.; Lacey, M.P. (**1999**). A method for high-sensitivity peptide sequencing using postsource decay matrix-assisted laser desorption ionization mass spectrometry. *Proc Natl Acad Sci U S A* **96**(13): 7131-6.

Keough, T.; Lacey, M.P.; Fieno, A.M.*; et al.* (**2000a**). Tandem mass spectrometry methods for definitive protein identification in proteomics research. *Electrophoresis* **21**(11): 2252-65.

Keough, T.; Lacey, M.P.; Youngquist, R.S. (**2000b**). Derivatization procedures to facilitate de novo sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **14**(24): 2348-56.

Keough, T.; Lacey, M.P.; Strife, R.J. (**2001**). Atmospheric pressure matrix-assisted laser desorption/ionization ion trap mass spectrometry of sulfonic acid derivatized tryptic peptides. *Rapid Commun Mass Spectrom* **15**(23): 2227-39.

Keough, T.; Lacey, M.P.; Youngquist, R.S. (**2002**). Solid-phase derivatization of tryptic peptides for rapid protein identification by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **16**(11): 1003-15.

Keough, T.; Youngquist, R.S.; Lacey, M.P. (**2003**). Sulfonic acid derivatives for peptide sequencing by MALDI MS. *Anal Chem* **75**(7): 156A-65A.

Kidwell, D.A.; Ross, M.M.; R.J., C. (**1984**). Sequencing of peptides by secondary ion mass spectrometry. *J Am Chem Soc* **106**(7): 2219-20.

Kimmel, J.R. **(1967)**. Guanidination of proteins. ed. Hirs, C. H. *Methods in enzymology*. San Diego CA, Academic Press. **11:** 584-89.

Klein, C.; Garcia-Rizo, C.; Bisle, B.*; et al.* (**2005**). The membrane proteome of Halobacterium salinarum. *Proteomics* **5**(1): 180-97.

Koller, A.; Washburn, M.P.; Lange, B.M.*; et al.* (**2002**). Proteomic survey of metabolic pathways in rice. *Proc Natl Acad Sci U S A* **99**(18): 11969-74.

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

Kyndt, J.A.; Fitch, J.C.; Meyer, T.E.*; et al.* (**2005**). Thermochromatium tepidum photoactive yellow protein/bacteriophytochrome/diguanylate cyclase: characterization of the PYP domain. *Biochemistry* **44**(12): 4755-64.

Lacey, M.P.; Keough, T.; Fieno, A.M.*; et al.* (**2000**). Definitive identification of proteins from bacteria having unsequenced genomes. *Proceedings of the 48th Conference on Mass Spectrometry and Allied Topics, Long Beach, CA.*

Lagerwerf, F.M.; van de Weert, M.; Heerma, W.*; et al.* (**1996**). Identification of oxidized methionine in peptides. *Rapid Commun Mass Spectrom* **10**(15): 1905-10.

Lee, Y.H.; Han, H.; Chang, S.B.*; et al.* (**2004a**). Isotope-coded N-terminal sulfonation of peptides allows quantitative proteomic analysis with increased de novo peptide sequencing capability. *Rapid Commun Mass Spectrom* **18**(24): 3019-27.

Lee, Y.H.; Kim, M.S.; Choie, W.S.*; et al.* (**2004b**). Highly informative proteome analysis by combining improved N-terminal sulfonation for de novo peptide sequencing and online capillary reverse-phase liquid chromatography/tandem mass spectrometry. *Proteomics* **4**(6): 1684-94.

Lee, Y.H.; Shin, Y.W.; Ryu, S.*; et al.* (**2006**). Enrichment of N-terminal sulfonated peptides by a water-soluble fullerene derivative and its applications to highly efficient proteomics. *Anal Chim Acta* **556**(1): 140-4.

Liao, P.C.; Huang, Z.H.; Allison, J. (**1997**). Charge remote fragmentation of peptides following attachment of a fixed positive charge: A matrix-assisted laser desorptio/ionization postsource decay study *J Am Soc Mass Spectrom* **8**: 501-9.

Lin, M.; Campbell, J.M.; Mueller, D.R.*; et al.* (**2003**). Intact protein analysis by matrix-assisted laser desorption/ionization tandem time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **17**(16): 1809-14.

Lindh, I.; Hjelmqvist, L.; Bergman, T.*; et al.* (**2000**). De novo sequencing of proteolytic peptides by a combination of C-terminal derivatization and nano-electrospray/collision-induced dissociation mass spectrometry. *J Am Soc Mass Spectrom* **11**(8): 673-86.

Liska, A.J.; Shevchenko, A. (**2003a**). Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics* **3**(1): 19-28.

Liska, A.J.; Shevchenko, A. (**2003b**). Combining mass spectrometry with database interrogation strategies in proteomics. *Trends Anal. Chem.* **22**: 291-8.

Liska, A.J.; Popov, A.V.; Sunyaev, S.*; et al.* (**2004a**). Homology-based functional proteomics by mass spectrometry: application to the Xenopus microtubule-associated proteome. *Proteomics* **4**(9): 2707-21.

Liska, A.J.; Shevchenko, A.; Pick, U.*; et al.* (**2004b**). Enhanced photosynthesis and redox energy production contribute to salinity tolerance in Dunaliella as revealed by homology-based proteomics. *Plant Physiol* **136**(1): 2806-17.

Loboda, A.V.; Krutchinsky, A.N.; Bromirski, M.*; et al.* (**2000**). A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption/ionization source: design and performance. *Rapid Commun Mass Spectrom* **14**(12): 1047-57.

Loo, J.A.; Edmonds, C.G.; Smith, R.D. (**1993**). Tandem mass spectrometry of very large molecules. 2. Dissociation of multiply charged proline-containing proteins from electrospray ionization. *Anal Chem* **65**(4): 425-38.

Lopez, M.F. (**2000**). Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. *Electrophoresis* **21**(6): 1082-93.

Lu, B.; Chen, T. (**2004**). Algorithms for de novo peptide sequencing using tandem mass spectrometry. *Drug Discov Today: Biosilico* **2**(2): 85-90.

Lysak, M.A.; Dolezelova, M.; Horry, J.P.*; et al.* (**1999**). Flow cytometric analysis of nuclear DNA content in Musa. *Theor. Appl. Gen.* **98**(8): 1344-50.

Mackey, A.J.; Haystead, T.A.; Pearson, W.R. (**2002**). Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol Cell Proteomics* **1**(2): 139-47.

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

Marekov, L.N.; Steinert, P.M. (**2003**). Charge derivatization by 4-sulfophenyl isothiocyanate enhances peptide sequencing by post-source decay matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *J Mass Spectrom* **38**(4): 373-7.

Marriott, J. (**1980**). Bananas--physiology and biochemistry of storage and ripening for optimum quality. *Crit Rev Food Sci Nutr* **13**(1): 41-88.

Martin, D.B.; Eng, J.K.; Nesvizhskii, A.I.*; et al.* (**2005**). Investigation of neutral loss during collision-induced dissociation of peptide ions. *Anal Chem* **77**(15): 4870-82.

Mathesius, U.; Imin, N.; Chen, H.*; et al.* (**2002**). Evaluation of proteome reference maps for cross-species identification of proteins by peptide mass fingerprinting. *Proteomics* **2**(9): 1288-303.

Matis, M.; Zakelj-Mavric, M.; Peter-Katalinic, J. (**2005**). Mass spectrometry and database search in the analysis of proteins from the fungus Pleurotus ostreatus. *Proteomics* **5**(1): 67-75.

Matthiesen, R.; Trelle, M.B.; Hojrup, P.*; et al.* (**2005**). VEMS 3.0: algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J Proteome Res* **4**(6): 2338-47.

Mayfield, J.A.; Fiebig, A.; Johnstone, S.E.*; et al.* (**2001**). Gene families from the Arabidopsis thaliana pollen coat proteome. *Science* **292**(5526): 2482-5.

Medzihradszky, K.F.; Campbell, J.M.; Baldwin, M.A.*; et al.* (**2000**). The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal Chem* **72**(3): 552-8.

Mohan, D.; Pasa-Tolic, L.; Masselon, C.D.*; et al.* (**2003**). Integration of electrokinetic-based multidimensional separation/concentration platform with electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry for proteome analysis of Shewanella oneidensis. *Anal Chem* **75**(17): 4432-40.

Munchbach, M.; Quadroni, M.; Miotto, G.*; et al.* (**2000**). Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal Chem* **72**(17): 4047-57.

Nesvizhskii, A.I.; Keller, A.; Kolker, E.*; et al.* (**2003**). A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **75**(17): 4646-58.

Ng, W.V.; Kennedy, S.P.; Mahairas, G.G.*; et al.* (**2000**). Genome sequence of Halobacterium species NRC-1. *Proc Natl Acad Sci U S A* **97**(22): 12176-81.

Nielsen, M.L.; Savitski, M.M.; Zubarev, R.A. (**2005**). Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Mol Cell Proteomics* **4**(6): 835-45.

Nold, M.J.; Wesdemiotis, C.; Yalcin, T.*; et al.* (**1997**). Amide bond dissociation in protonated peptides. Structures of the N-terminal ionic and neutral fragments. *Int J Mass spectr ion process* **164**(1-2): 137-53.

Oehlers, L.P.; Perez, A.N.; Walter, R.B. (**2005**). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of 4-sulfophenyl isothiocyanate-derivatized peptides on AnchorChip sample supports using the sodium-tolerant matrix 2,4,6-trihydroxyacetophenone and diammonium citrate. *Rapid Commun Mass Spectrom* **19**(6): 752-8.

Paizs, B.; Suhai, S.; Hargittai, B.*; et al.* (**2002**). Ab initio and MS/MS studies on protonated peptides containing basic and acidic amino acid residues I. Solvated proton vs. salt-bridged structures and the cleavage of the terminal amide bond of protonated RD-NH2. *Int J Mass spectrom* **219**(1): 203-32.

Paizs, B.; Suhai, S.; Harrison, A.G. (**2003**). Experimental and theoretical investigation of the main fragmentation pathways of protonated H-Gly-Gly-Sar-OH and H-Gly-Sar-Sar-OH. *J Am Soc Mass Spectrom* **14**(12): 1454-69.

Paizs, B.; Suhai, S. (**2004**). Towards understanding the tandem mass spectra of protonated oligopeptides. 1: mechanism of amide bond cleavage. *J Am Soc Mass Spectrom* **15**(1): 103-13.

Paizs, B.; Suhai, S. (**2005**). Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* **24**(4): 508-48.

Pashkova, A.; Moskovets, E.; Karger, B.L. (**2004**). Coumarin tags for improved analysis of peptides by MALDI-TOF MS and MS/MS. 1. Enhancement in MALDI MS signal intensities. *Anal Chem* **76**(15): 4550-7.

Pashkova, A.; Chen, H.S.; Rejtar, T.*; et al.* (**2005**). Coumarin tags for analysis of peptides by MALDI-TOF MS and MS/MS. 2. Alexa Fluor 350 tag for increased peptide and protein Identification by LC-MALDI-TOF/TOF MS. *Anal Chem* **77**(7): 2085-96.

Peiren, N.; Vanrobaeys, F.; de Graaf, D.C.*; et al.* (**2005**). The protein composition of honeybee venom reconsidered by a proteomic approach. *Biochim Biophys Acta* **1752**(1): 1-5.

Peltier, J.B.; Friso, G.; Kalume, D.E.*; et al.* (**2000**). Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* **12**(3): 319-41.

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

Porubleva, L.; Vander Velden, K.; Kothari, S.*; et al.* (**2001**). The proteome of maize leaves: use of gene sequences and expressed sequence tag data for identification of proteins with peptide mass fingerprints. *Electrophoresis* **22**(9): 1724-38.

Qin, J.; Chait, B.T. (**1995**). Preferential fragmentation of protonated gas-phase ions adjacent to acidic amino acid residues. *J Am Chem Soc* **117**(19): 5411-12.

Qin, J.; Herring, C.J.; Zhang, X. (**1998**). De novo peptide sequencing in an ion trap mass spectrometer with 18O labeling. *Rapid Commun Mass Spectrom* **12**(5): 209-16.

Quadroni, M.; James, P. (**1999**). Proteomics and automation. *Electrophoresis* **20**(4-5): 664-77.

Raucci, G.; Gabrielli, M.; Novelli, S.*; et al.* (**2005**). CHASE, a charge-assisted sequencing algorithm for automated homology-based protein identifications with matrix-assisted laser desorption/ionization time-of-flight post-source decay fragmentation data. *J Mass Spectrom* **40**(4): 475-88.

Roth, K.D.; Huang, Z.H.; Sadagopan, N.*; et al.* (**1998**). Charge derivatization of peptides for analysis by mass spectrometry. *Mass Spectrom Rev* **17**(4): 255-74.

Sadagopan, N.; Watson, J.T. (**2000**). Investigation of the tris(trimethoxyphenyl)phosphonium acetyl charged derivatives of peptides by electrospray ionization mass spectrometry and tandem mass spectrometry. *J Am Soc Mass Spectrom* **11**(2): 107-19.

Samyn, B.; Debyser, G.; Sergeant, K.*; et al.* (**2004**). A case study of de novo sequence analysis of N-sulfonated peptides by MALDI TOF/TOF mass spectrometry. *J Am Soc Mass Spectrom* **15**(12): 1838-52.

Savitski, M.M.; Nielsen, M.L.; Kjeldsen, F.*; et al.* (**2005**). Proteomics-grade de novo sequencing approach. *J Proteome Res* **4**(6): 2348-54.

Sergeant, K.; Samyn, B.; Debyser, G.*; et al.* (**2005**). De novo sequence analysis of N-terminal sulfonated peptides after in-gel guanidination. *Proteomics* **5**(9): 2369-80.

Shen, T.L.; Huang, Z.H.; Laivenieks, M.*; et al.* (**1999**). Evaluation of charge derivatization of a proteolytic protein digest for improved mass spectrometric analysis: de novo sequencing by matrix-assisted laser desorption/ionization post-source decay mass spectrometry. *J Mass Spectrom* **34**(11): 1154-65.

Shevchenko, A.; Chernushevich, I.; Ens, W.*; et al.* (**1997**). Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* **11**(9): 1015-24.

Shevchenko, A.; Loboda, A.; Shevchenko, A.*; et al.* (**2000**). MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research. *Anal Chem* **72**(9): 2132-41.

Shevchenko, A.; Sunyaev, S.; Loboda, A.*; et al.* (**2001**). Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem* **73**(9): 1917-26.

Shevchenko, A.; de Sousa, M.M.; Waridel, P.*; et al.* (**2005**). Sequence similarity-based proteomics in insects: characterization of the larvae venom of the Brazilian moth Cerodirphia speciosa. *J Proteome Res* **4**(3): 862-9.

Shui, W.; Liu, Y.; Fan, H.*; et al.* (**2005**). Enhancing TOF/TOF-based de novo sequencing capability for high throughput protein identification with amino acid-coded mass tagging. *J Proteome Res* **4**(1): 83-90.

Shütz, F.; Kapp, E.A.; Simpson, R.J.*; et al.* (**2003**). Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem Soc Transact* **31**: 1479-83.

Somogyi, A.; Wysocki, V.H.; Mayer, I. (**1994**). The effect of protonation site on bond strengths in simple peptides: Application of ab initio and modified neglect of differential overlap bond orders and modified neglect of differential overlap energy partitioning. *J Am Soc Mass Spectrom* **5**(8): 704.

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

Sprenger, W.W.; Hoff, W.D.; Armitage, J.P.*; et al.* (**1993**). The eubacterium Ectothiorhodospira halophila is negatively phototactic, with a wavelength dependence that fits the absorption spectrum of the photoactive yellow protein. *J Bacteriol* **175**(10): 3096-104.

Strahler, J.R.; Smelyanskiy, Y.; Lavine, G.*; et al.* (**1997**). Development of methods for the charge-derivatization of peptides in polyacrylamide gels and membranes for their direct analysis using matrix-assisted laser desorption-ionization mass spectrometry *Int J Mass Spectrom Ion Proc* **169/170**: 111-26.

Strosse, H.; Schoofs, H.; Panis, B.*; et al.* (**2006**). Development of embryogenic cell suspensions from shoot meristematic tissue in bananas and plantains (Musa spp.). *Plant Sci.* **170**(1): 104-12.

Stults, J.T.; Lai, J.; McCune, S.*; et al.* (**1993**). Simplification of high-energy collision spectra of peptides by amino-terminal derivatization. *Anal Chem* **65**(13): 1703-8.

Suckau, D.; Resemann, A. (**2003a**). T3-sequencing: targeted characterization of the N- and C-termini of undigested proteins by mass spectrometry. *Anal Chem* **75**(21): 5817-24.

Suckau, D.; Resemann, A.; Schuerenberg, M.*; et al.* (**2003b**). A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Anal Bioanal Chem* **376**(7): 952-65.

Summerfield, S.G.; Steen, H.; O'Malley, M.*; et al.* (**1999**). Phenyl thiocarbamoyl and related derivatives of peptides: Edman chemistry in the gas phase. . *Int J Mass Spectrom* **188**(1-2): 95-103.

Sumpton, D.P.; Thomas, J.R.; Thomas-Oates, J. (**2003**). Assessment of CID fragmentation behaviour of guanidinated, N-sulfonated tryptic peptides in MALDI Q-oTOF and MALDI TOF-TOF instrumentation. *16th International Mass Spectrometry Conference, Edinburgh*.

Sunyaev, S.; Liska, A.J.; Golod, A.*; et al.* (**2003**). MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* **75**(6): 1307-15.

Swennen, R.; Rosales, F. **(1994)**. Bananas. ed. Arntzen, C. *Encyclopedia of Agricultural Science*. New York, Academic Press. **1:** 215-32.

Tabb, D.L.; Smith, L.L.; Breci, L.A.*; et al.* (**2003**). Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem* **75**(5): 1155-63.

Tabb, D.L.; Huang, Y.; Wysocki, V.H.*; et al.* (**2004**). Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal Chem* **76**(5): 1243-8.

Taylor, J.A.; Johnson, R.S. (**2001**). Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem* **73**(11): 2594-604.

Tebbe, A.; Klein, C.; Bisle, B.*; et al.* (**2005**). Analysis of the cytosolic proteome of Halobacterium salinarum and its implication for genome annotation. *Proteomics* **5**(1): 168-79.

Tsaprailis, G.; Nair, H.; Somogyi, A.*; et al.* (**1999**). Influence of secondary structure on the fragmentation of protonated peptides. *J Am Chem Soc* **121**(22): 5142-54.

Tsaprailis, G.; Somogyi, A.; Nikolaev, E.N.*; et al.* (**2000**). Refining the model for selective cleavage at acidic residues in arginine-containing protonated peptides. *Int. J. Mass Spectrom.* **195/196**: 467-79.

Tsaprailis, G.; Nair, H.; Zhong, W.*; et al.* (**2004**). A mechanistic investigation of the enhanced cleavage at histidine in the gas-phase dissociation of protonated peptides. *Anal Chem* **76**(7): 2083-94.

Vaisar, T.; Urban, J. (**1996**). Probing the proline effect in CID of protonated peptides. *J Mass Spectrom* **31**(10): 1185-7.

van Wijk, K.J. (**2001**). Challenges and prospects of plant proteomics. *Plant Physiol* **126**(2): 501-8.

Vanrobaeys, F.; Devreese, B.; Lecocq, E.*; et al.* (**2003**). Proteomics of the dissimilatory iron-reducing bacterium Shewanella oneidensis MR-1, using a matrix-assisted laser desorption/ionization-tandem-time of flight mass spectrometer. *Proteomics* **3**(11): 2249-57.

VerBerkmoes, N.C.; Bundy, J.L.; Hauser, L.*; et al.* (**2002**). Integrating 'top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of Shewanella oneidensis. *J Proteome Res* **1**(3): 239-52.

Wagner, D.S.; Salari, A.; Gage, D.A.*; et al.* (**1991**). Derivatization of peptides to enhance ionization efficiency and control fragmentation during analysis by fast atom bombardment tandem mass spectrometry. *Biol Mass Spectrom* **20**(7): 419-25.

Wait, R.; Miller, I.; Eberini, I.*; et al.* (**2002**). Strategies for proteomics with incompletely characterized genomes: the proteome of Bos taurus serum. *Electrophoresis* **23**(19): 3418-27.

Walker, A.K.; Andrews, P.C. (**2003**). The effects of MALDI matrix and CID gas on the fragmentation efficiencies of peptides and proteins up to ~12,000 Da in a MALDI TOF/TOF mass spectrometer *Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics, Montreal*.

Wang, D.; Kalb, S.R.; Cotter, R.J. (**2004**). Improved procedures for N-terminal sulfonation of peptides for matrix-assisted laser desorption/ionization post-source decay peptide sequencing. *Rapid Commun Mass Spectrom* **18**(1): 96-102.

Wang, R.; Chait, B.T. (**1996**). Posttranslational modifications analyzed by automated protein ladder sequencing. *Methods Mol Biol* **61**: 161-70.

Wasinger, V.C.; Cordwell, S.J.; Cerpa-Poljak, A.*; et al.* (**1995**). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* **16**(7): 1090-4.

Wattenberg, A.; Organ, A.J.; Schneider, K.*; et al.* (**2002**). Sequence dependent fragmentation of peptides generated by MALDI quadrupole time-of-flight (MALDI Q-TOF) mass spectrometry and its implications for protein identification. *J Am Soc Mass Spectrom* **13**(7): 772-83.

Wilkins, M.R.; Williams, K.L. (**1997**). Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J Theor Biol* **186**(1): 7-15.

Wilkins, M.R.; Gasteiger, E.; Wheeler, C.H.*; et al.* (**1998**). Multiple parameter cross-species protein identification using MultiIdent--a world-wide web accessible tool. *Electrophoresis* **19**(18): 3199-206.

Wilm, M.; Shevchenko, A.; Houthaeve, T.*; et al.* (**1996**). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**(6564): 466-9.

Wysocki, V.H.; Tsaprailis, G.; Smith, L.L.*; et al.* (**2000**). Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* **35**(12): 1399-406.

Yalcin, T.; Khouw, C.; Csizmadia, I.G.*; et al.* (**1995**). Why are B ions stable species in peptide spectra? *J Am Soc Mass Spectrom* **6**(12): 1165-74.

Yamaguchi, M.; Nakazawa, T.; Kuyama, H.*; et al.* (**2005**). High-throughput method for N-terminal sequencing of proteins by MALDI mass spectrometry. *Anal Chem* **77**(2): 645-51.

Yan, S.P.; Zhang, Q.Y.; Tang, Z.C.*; et al.* (**2005**). Comparative proteomic analysis provides new insights into chilling stress responses in rice. *Mol Cell Proteomics*.

Yang, X.; Wu, H.; Kobayashi, T.*; et al.* (**2004**). Enhanced ionization of phosphorylated peptides during MALDI TOF mass spectrometry. *Anal Chem* **76**(5): 1532-36.

Yergey, A.L.; Coorssen, J.R.; Backlund, P.S., Jr.*; et al.* (**2002**). De novo sequencing of peptides using MALDI/TOF-TOF. *J Am Soc Mass Spectrom* **13**(7): 784-91.

Yu, W.; Vath, J.E.; Huberty, M.C.*; et al.* (**1993**). Identification of the facile gas-phase cleavage of the Asp-Pro and Asp-Xxx peptide bonds in matrix-assisted laser desorption time-of-flight mass spectrometry. *Anal Chem* **65**(21): 3015-23.

Zhang, W.; Krutchinsky, A.N.; Chait, B.T. (**2003**). "De novo" peptide sequencing by MALDI-quadrupole-ion trap mass spectrometry: a preliminary study. *J Am Soc Mass Spectrom* **14**(9): 1012-21.

Zhu, W.; Reich, C.I.; Olsen, G.J.*; et al.* (**2004**). Shotgun proteomics of Methanococcus jannaschii and insights into methanogenesis. *J Proteome Res* **3**(3): 538-48.

PART 4

CONCLUSIONS AND FUTURE PERSPECTIVES

4.1. Introduction

Proteomics research entails the global characterization of proteins expressed in cells under defined conditions. Because of the wide dynamic range of expressed proteins and the variability of the gene products, monitoring of protein expression profiles remains a challenging task. Current techniques are unable to explore all of the phenomena that occur at the protein level in-depth. In the work presented here, research was done to develop new methods for the determination of the C-terminal sequence of proteins and for the identification of proteins isolated from organisms whose genomic sequence is not known.

4.2. C-terminal sequence analysis

In contrast to the analysis of phosphorylation and glycosylation sites, relatively little attention has been paid to the development of approaches for the systematic analysis of proteolytic processing events. In Part 2, a new mass spectrometry-based strategy that allows the identification of the C-terminal sequence of proteins was presented. The method can be directly applied on proteins, cleaved with CNBr, purified either by SDS-PAGE, 2D-PAGE, or in solution, and therefore eliminates the specific isolation of the C-terminal peptide. Using *Shewanella oneidensis* as a model system, it was demonstrated that this approach can be used for C-terminal sequence analysis at the sensitivity level of most proteomic studies. In collaboration with Professor Faro (University of Coimbra, Portugal) the method was applied to study the C-terminal proteolytic processing of procardosin A, an aspartic acid protease isolated from *Cynara cardunculus*. Furthermore, a spiking experiment was performed in which the cellular protein extract from *Shewanella oneidensis* was spiked with 1% cardosin A (w/w) and separated by one and two dimensional gel electrophoresis. This demonstrated that our method can be applied on relatively complex samples.

In its current form, the method is not suitable for a high-throughput proteomic approach and the use of the MALDI TOF/TOF instrument requires that the molecular weight of the C-terminal peptide is below 5 kDa. Statistical analysis of the *Shewanella* protein database indicated that about 50% the proteins, encoded by the 5177 ORFs, have a C-terminal fragment that is identifiable using this approach.

The low sensitivity of reflectron MS-analysis for peptides with a mass higher than 5 kDa requires that linear MS-analyses are performed. However, the resolution and accuracy of the MALDI TOF/TOF instrument in the linear mode is too low to distinguish amino acids that differ by only a few Thomson units, Ile/Leu/Asn/Asp and Lys/Gln/Glu/Met. The use of a Fourier transform mass spectrometer, an instrument with a higher mass accuracy, will allow the unambiguous determination of these amino acids (Li *et al*, 1994; Marshall *et al*, 1998). Therefore, the analysis of the C-terminal ladders with such an instrument will likely improve the applicability of our method. For C-terminal sequence determinations of gel-separated proteins, the low extraction efficiency of large peptides further decreases the sensitivity of our approach. The extraction procedure that was applied in the work presented here is optimized for the extraction of tryptic peptides. However, the extraction yield will drop when larger polypeptides or proteins must be extracted. Recently, a new method for extraction of proteins up to 66.3 kDa (serum albumin) has been presented; incubation of gel spots in alkaline solutions resulted in the recovery of more then 50% of the proteins (Jin *et al*, 2005). Unfortunately, incubation in alkaline conditions will shift the equilibrium between homoserine lactone and its open form to the homoserine, and therefore cannot be applied in our method.

Another approach to solve the 5kDa limit is by using a cleavage method that generates smaller peptides. These can be generated by performing multiple cleavage steps or by performing a cleavage at the N- or C-terminus of a more abundant amino acid. Cleavage after more than one residue in a single experiment is another possibility to generate smaller C-terminal peptides. Irrespective of the cleavage method applied, the specific truncation of the original C-terminal peptide by carboxypeptidases must be preserved to allow enzymatic C-terminal sequence analysis. We are currently developing a new cleavage method, based on a protocol described in 1994 (Huang *et al*, 1994), that results in the cleavage of both tryptophanyl and methionyl peptide bonds. During cleavage of bonds C-terminal to tryptophan, the latter is converted to a Cγ-O-spirolactone derivative. This amino acid derivative has the same core structure as the homoserine lactone (Figure 4.1), a structure also observed after cleavage of proteins with BNPS-skatole (Crimmins *et al*, 1990; Rahali *et al*, 1999). The similarity between these two structures provides a reasonable argument that the activity of the used carboxypeptidases will be comparable on peptides ending on either amino acid derivative, an argument that is supported by results from initial experiments.

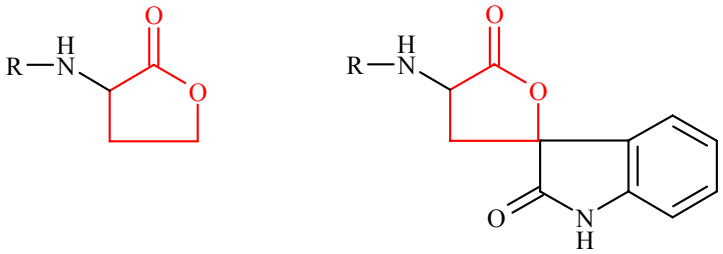


Figure 4.1. The reaction products after cleavage using our new chemical cleavage method. The homoserine lactone core (left), a structure also present in the tryptophan spirolactone derivative (right), is indicated in red.

Most enzymatic and chemical cleavage methods require that disulfide bonds in the proteins are reduced and the resulting sulfhydryl groups alkylated. These steps involve the use of concentrated salt solutions to denature the protein, which must be removed prior to analysis. Performing reduction, alkylation and desalting often results in sample loss and increases the risk for sample contamination. Alternatively, cysteine bridges can be broken by oxidation of disulfide bonds. Most often this is accomplished by incubation of the protein in performic acid (HCO_3H), resulting in the complete oxidation of the cysteines to cysteic acid residues. The new chemical cleavage method is an oxidative reaction during which it was observed that cystines are also completely oxidized to cysteic acid residues, eliminating the preliminar modification of the protein. The use of this cleavage protocol for other applications will be studied.

Independent of the applied cleavage method and of the type of mass spectrometer used for the analysis, automation of our method for C-terminal sequence analysis is imperative. The method, as described here, is labor-intensive and parallel sample handling of large batches is only possible in automated settings. Because highly toxic and corrosive chemicals must be manipulated, automation of this method is expected to be more troublesome than the automation described in Part 3.3.3.

4.3. *De novo* sequence analysis

Because protein identifications rely on non-error tolerant matches with sequence databases, high-throughput proteomics is currently largely restricted to those species for which comprehensive sequence databases are available. The identification of proteins derived from organisms with unsequenced genomes mainly depends on homology searching, an error tolerant approach. Homology searching requires that the sequence of peptides/proteins is determined solely based on experimental results, so called *de novo* sequence determination. Tandem mass spectrometric *de novo* sequence determination is severely hindered by the limited knowledge of the fragmentation mechanisms of peptide ions and the difficult interpretation of the often complex spectra. Currently, N-terminal derivatization of tryptic peptides with a negatively charged sulfonic acid group is a very promising way to eliviate the prohibitively difficult *de novo* interpretation of fragmentation spectra (Samyn *et al*, 2004).

Here, a novel approach in which gel-separated proteins are guanidinated in-gel prior to enzymatic cleavage is reported (Part 3.2). In contrast to previously described techniques, this procedure allows the extracted tryptic peptides to be sulfonated without any further sample purification. This approach allowed us to determine de novo peptide sequences of up to 24 amino acid residues in length. Subsequently, the improved protocol was applied on proteomic studies of 2D-PAGE separated proteins from *Halorhodospira halophila*, an extremophilic eubacterium, and banana (Musa spp.), organisms with an unsequenced genome (resp. Part 3.3.1 & 3.3.2). Using three different homology-based search algorithms, we were able to identify proteins from these organisms using sub-picomole quantities of protein. During these studies isoforms were identified for different proteins, and we characterized a novel PTM.

The automation of our improved protocol allowed to perform N-terminal sulfonation in a high-throughput fashion (Part 3.3.3). Furthermore, the data presented here are the first examples of the use of N-terminal sulfonation, after gel-separation of proteins, in medium to large-scale proteomic studies. However, only improvements in the front-end sample handling are reported. To attain a procedure that has a throughput comparable to non-error tolerant proteomic approaches, development of suitable algorithms for spectral interpretation and expedition of the database searches are imperative.

Recently, several publications indicated that compiling results from fully automated protein identification studies requires the careful judging of the certainty of identification (Ulintz *et al*, 2006; Wilkins *et al*, 2006). In contrast to data sets in transcriptomics, very few MS-based proteomics data sets are available to the general public and thus amendable for independent reevaluation (Prince *et al*, 2004). Nevertheless, it has been demonstrated that the public availability of transcriptomics data sets has resulted in reanalysis and new conclusions. For proteomics data sets, the need to reevaluate spectra is apparent by the fact that only a fraction of the fragmentation spectra in high-throughput MS experiments are satisfactorily interpreted. For example, in a large-scale analysis of the yeast proteome only 17% of 162,000 MS/MS spectra could be interpreted (Peng *et al*, 2003). In this study no distinction was made between high and low quality fragmentation spectra, a differentiation that was made by the group of K. Resing. This however, allowed to identify only 50% of the peptides using either SEQUEST or Mascot (Resing *et al*, 2004). An important explanation for these low efficiencies is that high-throughput platforms use non-error tolerant identification approaches. Consequently, peptides resulting from missed cleavages, unknown isoforms, adduct formation, common modifications (such as acetylation or methylation), single nucleotide polymorphisms and proteins erroneously represented in the databases, are not matched

(Resing *et al*, 2005). Decreasing the mass accuracy of the database search by allowing a larger mass window for the precursor, or by allowing more possible modifications, only partially resolves this problem. Furthermore, the possibility of false positive peptide identifications increases. As demonstrated, the use of our method for *de novo* sequence determination can alleviate some of these problems. Therefore, if the speed of data processing can be augmented, MDLC separation of sulfonated peptides (Lee *et al*, 2004; Flensburg *et al*, 2005), followed by MS/MS and homology-based database searches may be an attractive alternative for high-throughput identification of proteins isolated from eukaryotes.

4.4. Samenvatting en conclusies.

De grootste uitdaging binnen de levenswetenschappen vandaag is de volledige karakterisering van alle processen die een cel of weefsel tot een functioneel geheel maken. Voor moleculair biologische disciplines kon deze stap maar gezet worden na technische ontwikkelingen en verbeteringen in computergestuurde analyses. Deze alternatieve aanpak werd het eerst waargenomen in de sequentieanalyse van genomen. Het genoom van de eerste virussen werd al aan het einde van de jaren 70 volledig gesequeneerd (Fiers *et al*, 1978). Toch duurde het nog bijna 20 jaar voor het eerste genoom van een levend organisme werd bepaald, *Haemophilus influenza* (Fleischmann *et al*, 1995). Sindsdien werden meer dan 300 genomen gesequeneerd, met als belangrijkste mijlpaal de volledige bepaling van het humane genoom. Ondanks initiële verwachtingen bleek dat met de genoomsequentie van een organisme slechts een eerste indruk kan worden bekomen van de cellulaire complexiteit. De inherente stabiliteit van het genoom van een organisme laat niet toe om de dynamische en omgevingsafhankelijke cellulaire processen te beschrijven en geeft enkel het cellulaire potentieel weer. Een tijds- en omgevingsafhankelijke dimensie werd toegevoegd door de analyse van de functionele complementen van het genoom, zogenaamde 'functionele genoomanalyse'.

De implementatie van microarrayanalyses voor de studie van het transcriptoom leidde tot snelle kwantitatieve analyses van alle geactiveerde genen en geeft aldus een eerste inzicht in de regularisatie van de cellulaire processen. De activatie van genen is echter slechts een eerste controlemechanisme; verschillende regulariserende processen spelen zich af op eiwitniveau, na translatie van het mRNA in aminozuursequenties. De volledige functionele analyse van een organisme kan dan ook enkel worden gerealiseerd op eiwitniveau. Deze analyses worden gegroepeerd onder de term proteoomanalyse, gedefinieerd als de analyse van het eiwitcomplement aan een genoom van een organisme of cel op een bepaald tijdstip onder welbepaalde omstandigheden (Wilkins *et al*, 1996). De facto werden de eerste holistische eiwitanalyses al uitgevoerd in 1975 met de ontwikkeling van de 2D-gelelektroforesetechniek, lang vóór de term proteoom werd gelanceerd (O'Farrell, 1975). Grootschaligheid, een kenmerk van de huidige aanpak, werd maar mogelijk na de ontwikkeling van verbeterde massaspectrometrische analysemethodes, namelijk elektrosprayionisatie (ESI) (Fenn *et al*, 1989) en matrix geassisteerde laser desorptie/ionisatie (MALDI-MS) (Karas *et al*, 1985; Tanaka *et al*, 1988), ontwikkelingen die in 2002 werden gehonoreerd met de toekenning van de Nobelprijs Scheikunde aan Fenn en Tanaka. Een tweede vereiste om grootschalige eiwitanalyses toe te laten was de ontwikkeling van algoritmes waarmee massaspectra kunnen gecorreleerd worden met de beschikbare gegevens in databanken. Voor het correleren van de peptidenmassa's met databankgegevens werden gelijktijdig door verschillende onderzoeksgroepen verschillende algoritmes voorgesteld (Henzel *et al*, 1993; James *et al*, 1993; Mann *et al*, 1993; Pappin *et al*, 1993; Yates *et al*, 1993). Om eiwitten met hogere specificiteit te identificeren kan gebruikt worden gemaakt van tandem-massaspectrometrische technieken. Software voor de analyse van dergelijke spectra is meestal gebaseerd op een algoritme dat in 1994 werd beschreven (Eng *et al*, 1994). Vandaag is de analyse van het proteoom van een organisme mogelijk met volledig geautomatiseerde methodes, zoals de koppeling van meerdimensionale vloeistofchromatografie met massaspectrometrie (Washburn *et al*, 2001).

De dynamiek die wordt waargenomen op eiwitniveau is te wijten aan strikt gecontroleerde processen die plaats vinden tijdens of na de translatie van nucleotide- naar aminozuursequentie. Deze processen leiden ofwel tot het vormen of breken van covalente bindingen, bijvoorbeeld posttranslationele modificatie van aminozuurzijketens, N- of C-

terminale verkorting van de eiwitketen (proteolytische splitsing) en gecontroleerde eiwitdegradatie, of in het vormen van zwakkere bindingen zoals waterstofbruggen en hydrofobe interacties bij het vormen van eiwitcomplexen. De regularisatie, het resultaat en de impact van deze processen kan niet a-priori uit genomische data worden afgeleid. Dientengevolge kan enkel de analyse van de moleculen die het feitelijke fenotype van een organisme bepalen, de eiwitten, leiden tot de beschrijving van deze processen (Pandey *et al*, 2000).

C-terminale sequentieanalyse

Voor de studie van de meest abundante posttranslationele modificaties (PTM's), zoals fosforylatie en glycosylatie, werden talrijke technieken ontwikkeld, dit in tegenstelling tot andere belangrijke PTM's waaraan veel minder aandacht werd besteed. C-terminale proteolytische splitsing is belangrijk in een aantal cellulaire processen en er bestaan verschillende ziektebeelden die gerelateerd zijn met een verkeerde splitsing van eiwitten. Het functioneel belang van de C-terminus van een eiwit, meestal buiten het globulaire deel van het eiwit gelegen en dus ideaal gesitueerd voor interactie met andere eiwitten, werd in verschillende recente publicaties aangetoond (Fanning *et al*, 1999; Chung *et al*, 2002; Chung *et al*, 2003). Toch bestonden bij aanvang van het hier beschreven werk geen technieken die C-terminale sequentieanalyse van eiwitten toelaten met een gevoeligheid vergelijkbaar met die van andere technieken in de studie van het proteoom. In deel 2.2 wordt een nieuwe techniek voor C-terminale sequentieanalyse beschreven. Hierin wordt de chemische splitsing van eiwitten met cyanogeen bromide (CNBr) gecombineerd met carboxypeptidase digestie van de resulterende peptidenmengsels (Samyn *et al*, 2005). Incubatie van een eiwit met CNBr leidt tot splitsing van polypeptideketens C-terminaal van methionine; tijdens deze reactie wordt methionine omgezet tot een homoserinelacton-derivaat dat in evenwicht is met zijn open vorm (Gross *et al*, 1962). Er werd vastgesteld dat de gebruikte carboxypeptidasen dit aminozuurderivaat, dat C-terminaal voorkomt op alle, behalve het originele C-terminale peptide, niet kunnen afsplitsen. Aldus wordt enkel het C-terminale peptide stapsgewijs afgebroken. In de massaspectra wordt deze degradatie waargenomen als een opeenvolging van pieken, de zogenaamde 'sequentieladders'. De C-terminale sequentie kan vervolgens worden bepaald door de massaverschillen tussen opeenvolgende pieken in de sequentieladders te berekenen. De gevoeligheid van deze methode werd aangetoond met de bepaling van de C-terminale sequentie van een aantal, met 2D-gelelektroforese gescheiden, eiwitten geïsoleerd uit *Shewanella oneidensis*.

De studie van de auto-activatie van cardosine A, een aspartaatprotease geïsoleerd uit *Cynara cardunculus*, toont het belang aan voor de studie van biologische stalen in het algemeen (Castanheira *et al*, 2005). De methode, zoals hier beschreven, heeft echter een aantal beperkingen. Zo kan, met de massaspectrometer die in deze studie werd gebruikt, enkel de C-terminale sequentie worden bepaald van eiwitten die een C-terminaal peptide bezitten met een massa lager dan 5 kDa (na CNBr-splitsing). In de toekomst wordt dan ook het gebruik van een Fourier-transform massaspectrometer in het vooruitzicht gesteld, een toestel dat accuratere massabepalingen toelaat op grotere fragmenten. Toepassen van een nieuwe splitsingsmethode en automatisatie moet resulteren in een methode die toelaat om de C-terminale sequentie van een groot aantal stalen snel en met hoge gevoeligheid te bepalen.

De novo sequentieanalyse

Met de meest courant gebruikte technieken is het succes van eiwitidentificatie sterk afhankelijk van het voorkomen van het eiwit in sequentiedatabanken. Hierdoor zijn de meeste high-throughput studies beperkt tot een aantal goed gekarakteriseerde organismen, de zogenaamde modelorganismen. Voor de identificatie van eiwitten uit niet-modelorganismen wordt gebruik gemaakt van de sequentiehomologie tussen experimentele gegevens en gekende eiwitten. In principe kunnen verschillende eigenschappen van eiwitten worden gebruikt voor zogenaamde 'cross-species' eiwitidentificatie (Wilkins *et al*, 1997). In praktijk blijken vooral de aminozuursamenstelling, de massa en de sequentie van een eiwit geschikt te zijn. Sequentiebepaling van eiwitten of peptiden, zonder afhankelijk te zijn van databanken, kan op verschillende manieren gebeuren. In het verleden was de meest gebruikte methode N-terminale sequentiebepaling met behulp van de Edmandegradatie. Tegenwoordig wordt fragmentatie van peptiden met tandem massaspectrometrische technieken verkozen vanwege de snelheid en gevoeligheid van de techniek. Omdat de fragmentatiemechanismen van peptiden nog maar gedeeltelijk gekend zijn is massaspectrometrische *de novo* sequentieanalyse echter niet eenvoudig (Paizs *et al*, 2005). Ook het gebruik van algorithmes voor in-silico *de novo* sequentieanalyse resulteren vaak in de bepaling van foute sequenties (Liska *et al*, 2003). Zelfs indien de volledige sequentie van een peptide kan worden bepaald is dit geen garantie dat een onbekend eiwit kan worden geïdentificeerd. Courant gebruikte algoritmes voor homologie-vergelijking, zoals FASTA en BLAST, presteren slecht indien slechts een korte sequentie wordt ingegeven. Om hieraan tegemoet te komen werden een aantal programma's ontwikkeld die geoptimaliseerd zijn voor het gebruik van massaspectrometrisch bepaalde sequenties voor 'cross-species' eiwitidentificatie.

In het verleden werden een aantal derivatisatiemethodes ontwikkeld om de bepaling van aminozuursequenties op basis van tandem MS spectra te vergemakkelijken. In een eerste groep van methodes worden peptiden gederivatiseerd met een gefixeerde lading (Roth *et al*, 1998). De basis voor een efficiëntere aanpak werd in 1992 gelegd toen fragmentatiespectra van cysteïnezuur-bevattende peptiden werden vergeleken met fragmentatiespectra van dezelfde peptiden die niet geoxideerd werden (Burlet *et al*, 1992). In de fragmentatiespectra van de geoxideerde peptiden domineerde één type van ionen (y-ionen) en het berekenen van de massaverschillen tussen deze ionen liet eenvoudige sequentiebepaling van deze peptiden toe. Dit gegeven werd door de onderzoeksgroep van Keough verder uitgewerkt tot een methode, de zogenaamde N-terminale sulfonylatie van peptiden (Keough *et al*, 1999). Bij het uitoeren van deze methode treedt tevens derivatisatie van de ε-aminogroep van lysine op en daarom werd de methode in initiële publicaties enkel toegepast op peptiden die geen lysine bevatten. Door het uitvoeren van guanidilatie, waarbij de ε-aminogroep van lysine wordt gederivatiseerd tot een guanidino-groep, kon deze methode meer algemeen worden toegepast (Keough *et al*, 2000). Recent werden alternatieve reagentia voor N-terminale sulfonylatie voorgesteld die, in tegenstelling tot het originele reagens, ook kunnen worden gebruikt in waterige oplossingen (Gevaert *et al*, 2001; Keough *et al*, 2002; Pashkova *et al*, 2005).

De uitvoering van guanidilatie voor N-terminale sulfonylatie vereist normaal dat een ontzoutingsstap wordt uitgevoerd. Deze stap leidt echter tot staalverlies en is bovendien moeilijk automatiseerbaar. In ons werk wordt een verbeterde methode voorgesteld waarin deze ontzoutingsstap kan vermeden worden, wat eveneens toelaat on het protocol te automatiseren.

Het protocol voor N-terminale sulfonylatie werd vereenvoudigd door het uitvoeren van de guanidilatie vóór splitsing van het eiwit met trypsine, terwijl het eiwit geïmmobiliseerd blijft in het polyacrylamidegel. Hierdoor blijft het verlies van staal tijdens deze stap beperkt. Na guanidilatie wordt het staal verder behandeld zoals in standaard trypsine digestieprotocols. Na extractie van de resulterende peptiden kunnen deze worden gesulfonyleerd en geanalyseerd zonder bijkomende opzuiveringsstappen (Sergeant *et al*, 2005). Vervolgens werd dit protocol toegepast in de studie van eiwitten geïsoleerd uit twee verschillende organismen waarvan geen sequentie-informatie in databanken voorhanden is.

In het laatste deel van dit werk werd een vergelijking gemaakt tussen het uitvoeren van N-terminale sulfonylatie met het cyclisch anhydride van 2-sulfobenzeenzuur en het watercompatibele reagens 4-sulfofenyl-isothiocyanaat (SPITC) (Gevaert *et al*, 2001; Marekov *et al*, 2003). Hiertoe werden twee identieke stalen gederivatiseerd en geanalyseerd. Na derivatisatie met SPITC werden eiwitten geïdentificeerd in 24 van de 30 uitgesneden bandjes; na derivatisatie met het cyclisch anhydridereagens werden slechts 3 eiwitten geïdentificeerd. Hiermee werd aangetoond dat gebruik van een watercompatibel reagens een positieve invloed heeft op de resultaten.

In samenwerking met het 'Laboratoire de protéomique' van het CRPGL (Luxemburg) werd de automatisatie van het in-gel guanidilatieprotocol, gevolgd door sulfonylatie met SPITC, gerealiseerd. Na optimalisatie werd een eerste experiment uitgevoerd op 35 spotjes uit een 2D-gel van eiwitten geïsoleerd uit het meristeem van banaan. Hierbij werden in 28 van de 35 spotjes eiwitten geïdentificeerd (80%). In samenwerking met het Laboratorium voor Tropische Plantenteelt (KUL) zal de identificatie van eiwitten geïsoleerd uit banaan worden voortgezet. Hoewel de beschreven resultaten aantonen dat automatisatie van het in-gel guanidilatieprotocol mogelijk is, vormt dit slechts een eerste stap naar een high-throughput methode voor *de novo* sequentieanalyse en 'cross-species' identificatie van eiwitten.

References

Burlet, O.; Yang, C.; Gaskell, S.J. (**1992**). Influence of cysteine to cysteic acid oxidation on the collision-activated decomposition of protonated peptides: evidence for intraionic interactions. *J Am Soc Mass Spectrom* **3**(4): 337-44.

Castanheira, P.; Samyn, B.; Sergeant, K.*; et al.* (**2005**). Activation, proteolytic processing, and peptide specificity of recombinant cardosin A. *J Biol Chem* **280**(13): 13047-54.

Chung, J.J.; Shikano, S.; Hanyu, Y.*; et al.* (**2002**). Functional diversity of protein C-termini: more than zipcoding? *Trends Cell Biol* **12**(3): 146-50.

Chung, J.J.; Yang, H.; Li, M. (**2003**). Genome-wide analyses of carboxyl-terminal sequences. *Mol Cell Proteomics* **2**(3): 173-81.

Crimmins, D.L.; McCourt, D.W.; Thoma, R.S.*; et al.* (**1990**). In situ chemical cleavage of proteins immobilized to glass-fiber and polyvinylidenedifluoride membranes: cleavage at tryptophan residues with 2-(2'-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine to obtain internal amino acid sequence. *Anal Biochem* **187**(1): 27-38.

Eng, J.K.; McCormack, A.L.; Yates, J.R.I. (**1994**). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**(11): 976-89.

Fanning, A.S.; Anderson, J.M. (**1999**). PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. *J Clin Invest* **103**(6): 767-72.

Fenn, J.B.; Mann, M.; Meng, C.K.*; et al.* (**1989**). Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**(4926): 64-71.

Fiers, W.; Contreras, R.; Haegemann, G.*; et al.* (**1978**). Complete nucleotide sequence of SV40 DNA. *Nature* **273**(5658): 113-20.

Fleischmann, R.D.; Adams, M.D.; White, O.*; et al.* (**1995**). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**(5223): 496-512.

Flensburg, J.; Tangen, A.; Prieto, M.*; et al.* (**2005**). Chemically-assisted fragmentation combined with multi-dimensional liquid chromatography and matrix-assisted laser desorption/ionization post source decay, matrix-assisted laser desorption/ionization tandem time-of flight or matrix-assisted laser desorption/ionization tandem mass spectrometry for improved sequencing of tryptic peptides. *Eur J Mass Spectrom* **11**(2): 169-79.

Gevaert, K.; Demol, H.; Martens, L.*; et al.* (**2001**). Protein identification based on matrix assisted laser desorption/ionization-post source decay-mass spectrometry. *Electrophoresis* **22**(9): 1645-51.

Gross, E.; Witkop, B. (**1962**). Nonenzymatic cleavage of peptide bonds: the methionine residues in bovine pancreatic ribonuclease. *J Biol Chem* **237**: 1856-60.

Henzel, W.J.; Billeci, T.M.; Stults, J.T.*; et al.* (**1993**). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A* **90**(11): 5011-5.

Huang, S.S.; huang, J.S. (**1994**). Cleavage of both tryptophanyl and methionyl peptide bonds in proteins. *J Protein Chem* **13**(5): 450-51.

James, P.; Quadroni, M.; Carafoli, E.*; et al.* (**1993**). Protein identification by mass profile fingerprinting. *Biochem Biophys Res Commun* **195**(1): 58-64.

Jin, Y.; Manabe, T. (**2005**). High-efficiency protein extraction from polyacrylamide gels for molecular mass measurement by matrix-assisted laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* **26**(6): 1019-28.

Karas, M.; Bachmann, D.; Hillenkamp, F. (**1985**). Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Anal Chem* **57**(14): 2935-9.

Keough, T.; Youngquist, R.S.; Lacey, M.P. (**1999**). A method for high-sensitivity peptide sequencing using postsource decay matrix-assisted laser desorption ionization mass spectrometry. *Proc Natl Acad Sci U S A* **96**(13): 7131-6.

Keough, T.; Lacey, M.P.; Youngquist, R.S. (**2000**). Derivatization procedures to facilitate de novo sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **14**(24): 2348-56.

Keough, T.; Lacey, M.P.; Youngquist, R.S. (**2002**). Solid-phase derivatization of tryptic peptides for rapid protein identification by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **16**(11): 1003-15.

Lee, Y.H.; Kim, M.S.; Choie, W.S.*; et al.* (**2004**). Highly informative proteome analysis by combining improved N-terminal sulfonation for de novo peptide sequencing and online capillary reverse-phase liquid chromatography/tandem mass spectrometry. *Proteomics* **4**(6): 1684-94.

Li, Y.; McIver, R.T., Jr.; Hunter, R.L. (**1994**). High-accuracy molecular mass determination for peptides and proteins by Fourier transform mass spectrometry. *Anal Chem* **66**(13): 2077-83.

Liska, A.J.; Shevchenko, A. (**2003**). Combining mass spectrometry with database interrogation strategies in proteomics. *Trends Anal. Chem.* **22**: 291-8.

Mann, M.; Hojrup, P.; Roepstorff, P. (**1993**). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom* **22**(6): 338-45.

Marekov, L.N.; Steinert, P.M. (**2003**). Charge derivatization by 4-sulfophenyl isothiocyanate enhances peptide sequencing by post-source decay matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *J Mass Spectrom* **38**(4): 373-7.

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

O'Farrell, P.H. (**1975**). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250**(10): 4007-21.

Paizs, B.; Suhai, S. (**2005**). Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* **24**(4): 508-48.

Pandey, A.; Mann, M. (**2000**). Proteomics to study genes and genomes. *Nature* **405**(6788): 837-46.

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

Pashkova, A.; Chen, H.S.; Rejtar, T.*; et al.* (**2005**). Coumarin tags for analysis of peptides by MALDI-TOF MS and MS/MS. 2. Alexa Fluor 350 tag for increased peptide and protein Identification by LC-MALDI-TOF/TOF MS. *Anal Chem* **77**(7): 2085-96.

Peng, J.; Elias, J.E.; Thoreen, C.C.*; et al.* (**2003**). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2**(1): 43-50.

Prince, J.T.; Carlson, M.W.; Wang, R.*; et al.* (**2004**). The need for a public proteomics repository. *Nature Biotech* **22**(4): 471-2.

Rahali, V.; Gueguen, J. (**1999**). Chemical cleavage of bovine beta-lactoglobulin by BNPS-skatole for preparative purposes: comparative study of hydrolytic procedures and peptide characterization. *J Protein Chem* **18**(1): 1-12.

Resing, K.A.; Meyer-Arendt, K.; Mendoza, A.M.*; et al.* (**2004**). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem* **76**(13): 3556-68.

Resing, K.A.; Ahn, N.G. (**2005**). Proteomics strategies for protein identification. *FEBS Lett* **579**: 885-9.

Roth, K.D.; Huang, Z.H.; Sadagopan, N.*; et al.* (**1998**). Charge derivatization of peptides for analysis by mass spectrometry. *Mass Spectrom Rev* **17**(4): 255-74.

Samyn, B.; Debyser, G.; Sergeant, K.*; et al.* (**2004**). A case study of de novo sequence analysis of N-sulfonated peptides by MALDI TOF/TOF mass spectrometry. *J Am Soc Mass Spectrom* **15**(12): 1838-52.

Samyn, B.; Sergeant, K.; Castanheira, P.*; et al.* (**2005**). A new method for C-terminal sequence analysis in the proteomic era. *Nat Methods* **2**(3): 193-200.

Sergeant, K.; Samyn, B.; Debyser, G.*; et al.* (**2005**). De novo sequence analysis of N-terminal sulfonated peptides after in-gel guanidination. *Proteomics* **5**(9): 2369-80.

Tanaka, K.; Waki, H.; Ido, Y.*; et al.* (**1988**). Protein and polymer analyses up to m/z 100.000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **2**(2): 151-3.

Ulintz, P.J.; Zhu, J.; Qin, Z.S.*; et al.* (**2006**). Improved classification of mass spectrometry database search results using newer machine learning approaches. *Mol Cell Proteomics* **5**(3): 497-509.

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**(3): 242-7.

Wilkins, M.R.; Sanchez, J.C.; Gooley, A.A.*; et al.* (**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.

Wilkins, M.R.; Williams, K.L. (**1997**). Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J Theor Biol* **186**(1): 7-15.

Wilkins, M.R.; Appel, R.D.; Van Eyk, J.E.*; et al.* (**2006**). Guidelines for the next 10 years of proteomics. *Proteomics* **6**(1): 4-8.

Yates, J.R., 3rd; Speicher, S.; Griffin, P.R.*; et al.* (**1993**). Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem* **214**(2): 397-408.

APPENDIXES

APPENDIX I

DE NOVO DERIVED PEPTIDE SEQUENCES FROM *SHEWANELLA ONEIDENSIS* PROTEINS DISCUSSED IN PART 3.2

Spot[a]	Protein[b]	FASTS[c]	E-score[d]	De novo derived sequences
1	15640750	antioxidant, AhpC/Tsa family	3.70e-39	MEEFTKR GVVVIGVSIDSQFTHN GSFIIDKEGMVR AYDVEHPEAGVAFR HQVVNDI NVDEMIR
2	24373389	conserved hypothetical protein	1.8e-26	IIIDAYSIER IEDSYIR EITKGIR IINTAEVTIAEKYR
3	24376191	conserved hypothetical protein	1.5e-18	FVADGFGVEPR ESKTAEEAFAK IAKSGIFPISR
4	24373827	nucleoside diphosphate kinase	3.1e-40	EIIGATNPAQAAXXXR ADFAESIDENAAHGSDSIA EIAYFFSAEXXCPR TFSIIKPDAVAK
5	24374659	DNA-binding prot, H-NS family	6.30e-22	IDKIXXER NAKIEEIR YQIEVDGE SEFIEIITHGR
6	24372148	hypothetical protein SO0554	8.30e-06	VVKDIYPPR DIKSS
7	24371815	translation elongation factor Tu	7.70e-27	KIIDEGR TTTKTTCTREMFR MKVTIICPI DIDKXXIM TYGXXXKDFSQIDNA KTINTSHIEYDT
8	24372773	conserved hypothetical protein	5.1e-20	DAYDIK VYIPDXKPSS GKEEVVR YFSVYKVK
9	24372802	purine nucleoside phosphorylase	7.00e-09	NVEQVTDVR AICVVTVSDXXAR
10	24375179	universal stress protein family	5.90e-26	SGQIIWXXDFSETAAH TGISHFNHTNVAESVADGAVC
11	24373389	conserved hypothetical protein	1.2e-26	IIIDAYSIER IINTAEVTIAEKYR EITKGIR KQGAIXXFAI GTSDKIEDS
12	24372359	malate dehydrogenase	1.8e-38	FGMSIVR IFGVTTIDV(IR/VK) SDIFNINAGIVR AGVYDKNR DIIEKVAVTCPK
13	24371943	methylisocitrate lyase	1.8e-24	IKAAVDAR IASYIDXXSR TDAVAVEGIEAGIER TDPNFVIMAR
14	24373372	phage shock protein A	4.80e-31	QKTIIIR

				IKDEVSIIQEK VIAEKKEIHR VKGKAPTK IIIQEMXXXXVEVR
15	24374936	carbon storage regulator	4.40e-12	IGVNAPKEVWHR IQSE NKEVSVHREEI
16	24376191	conserved hypothetical protein	3.00e-19	ESKTAEEAFAK IAKSGI FVADGFGVEPR
17	24373949	ribosomal protein S1	7.20e-11	YRANDXI VIKYDR AFIXXXIVDVRPVR QIGED
18	24373949	ribosomal protein S1	4.00e-37	DTAHIEY QIGEDXXIEISKR VKGGFXXEINXXR RAVIESE YIINDRITGR AFIXXXIVDV VIKYDR ISIGIKQCK
19	24372295	chaperonin GroEL	1.40e-51	VEDAIHATR AAKEVVFXXDAR VVITKDNTTIIDGDGEQAQXXAR IADVEVANEDQKHGVV AAVEEGVV
20	24372709	chaperone protein DnaK	1.80e-47	IINEPTAAAIAYGIDKK IAGIEVKR FKDDEVQR AKIESIVEDIIIR VQEAVVDFFGKEPR KDPIAMQR
21	24376221	ATP synthase F1, alpha subunit	6.70e-37	KISGGIR AQIEHGVR EAYPGDVFYIHSR MQINS FAIAINDQR IEQFEVV EIAAFSQFAS
22	24376191	conserved hypothetical protein	6.10e-16	ESKTAEEAFAK IAKSGI TAEEAFA GDNSGTHIK
23	24374657	electr transf flavoprot, alpha subunit	5.30e-22	SITVSARPEIGNAGIIVSNR IDTAKVVTAAR TIDKV
24a	24373875	translation elongation factor P	6.70e-11	VGDKIEIDTR NAAIVK EVIYTE VVQKT
24b	24372295	chaperonin GroEL	5.30e-06	AAKEVVFGNDAR
25	24373389	conserved hypothetical protein	7.90e-27	EITKGIR IINTAEVTIAEKYR GWDKIED

				IIIDAYSIER
26	24375049	OmpA family protein	4.60e-07	IIVEKYG EYYKDIER IEAIV VKEVG
27	24375049	OmpA family protein	2.70e-16	IIVEKYGIS EYYKDIER RIEAIVTXXEKQ
28	24371854	DNA-RNA polymerase α-chain	2.40e-43	GFGHTIGNAIR IVDIEQVN IAYNVEAAR AKVTIE SITEIKDXXASR MQGSVTEFIKPR
29	24371854	DNA-RNA polymerase α-chain	3.10e-10	SITEIKD GFGHTIGN AKVTIE
30	24371854	DNA-RNA polymerase α-chain	2.30e-27	AKVTIEPIER SITEIKD MQGSVTEFIKPR GFGHVIGNAIR
31	24373198	translation elongation factor Ts	2.40e-35	RVEYIDGAK VTNFIR AISAAQVKEIR KTVGEFIK VTIEDIKAQFEEAR
32	24373496	succinyl-CoA synth, beta-subunit	2.00e-36	IKAMHD MNIHEYQAK AIID IIVTYXXDEKGQPVAK VTGDKEEIR CQVHAGGR EVGVKV
33	24373197	ribosomal protein S2	6.90e-14	DMIQAGVHF VIFVGTKR EIEKIEK IRIENEIIXXIR
34	24375430	serine prot, HtrA/DegQ/DegS fam	6.90e-35	YFFGXXAPQE NIIAQIAEHG KAGDIIV NIVXXKVXXSDEIR AIKSFQEIR QRVPDVFR
35	24375430	serine prot, HtrA/DegQ/DegS fam	3.00e-43	GKSMIYIVIR YFFGPNAPQEQVQER AIKSXXEIR NIIAQIAEH AGIKAGDIIVSVDGR
36	24372294	chaperonin GroES	8.80e-25	VGDVVIFNE GEVIAV VIVKR IEVESTSAGXXVI MNIRPIHDR IIEDGT IRHPIDR

37	24371821	ribosomal protein L7/L12	2.2e-16	GATGIGIXEAK KEIVEAGASVEIK VAVIKAIR
38	24373198	translation elongation factor Ts	4.80e-21	AISAAQVKEIR VAIVAKIGENINVR RVEYID
39	24372359	malate dehydrogenase	2.10e-39	AGVYDKNR NIIEKVAVTCPK FGMSIVR IFGVTTIDVIR SDIFNINAGIVR
40	24375415	ribosomal protein L9	7.10e-55	AVVANESNVKV IAADIAAA AGDEGKIFGSVGNR NYIIPQGK GVEIAKSEVR RAEIEAK DIADAVTAAGVEIAKSEVR
41	24372202	stringent starvation protein a	7.00e-20	SVMTIFSGADDIYSH ESFKASITE EIVIYESR IDTDWYSIVAR
42	24374593	molybd ABC transp, periplasmic	7.00e-28	IAVGDPDHVPAGQ QAIENIXXWK AFNQYIQ VVFE QAIENLNLWKTAE
43	24374881	conserved hypothetical protein	2.30e-12	DKHIR AHAVGEGQD GEDFINTPIFAK
44	24374593	3-oxoacyl-(acyl-carrier-prot) synth	9.80e-21	IKIDXXEIIDR GITHSAQ EIEAIR SGITHS VGPYIVPR EKGVKR
45a	24376219	ATP synthase F1, beta subunit	2.80e-17	YTIAGT DVIIFV YVIHR AHSGISV VAITGIT DEGRDVII
45b	24375475	aerobic respirat control prot ArcA	7.20e-53	SIVSXXGESYKIPR EINNIGIIFI AMIHFVEN MQNPHIIIVEDEA VNSAGNEVEEKISVE EIKPHDR
46	24376191	conserved hypothetical protein	2.10e-14	IAKSGI ESKTAEEAFAK FVADXXXXEPR
47	24373359	trigger factor	4.50e-37	KQHATFAAVER QDITGEVMQR QQAMQR YGAAIR MQVSVEAVQGIER

				NVAIEEQAVE
48	24374298	uracil phosphoribosyltransferase	5.40e-16	KVTVVPIIR EGDIS AGIGMMDGVIEHI
49	24374881	conserved hypothetical protein	1.00e-09	DKHIR GEDFINTGKFAK
50	24372520	fruct-bisphosphate aldol, class II	1.00e-14	IKEIHAR KVNIDTDIR KYIAEHXXEF YEAF
51	24373111	isocitrate dehydrogenase	1.10e-23	YENI DIGGTHGTTD KAVSAVIEEGDR TIEIIEKNR
52	24372333	iron(III) ABC transp, periplasmic	7.10e-18	QAFIVEPIIKR DQKITVYSYR NIYTAKDR
53	24373497	succinyl-CoA synth, alpha-subunit	2.40e-12	SVIIN VKIEETGVR SIADIGKAIR

[a] spot number according to the position on the 2D-PAGE (Figure 3.14)
[b] NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/)
[c] protein with lowest E-value in FASTS search result
[d] In FASTS, the E(N) value reports the number of times the score should be obtained by chance against a database of size N. For searches against the NCBI non-redundant protein databases N ≈ 2075116.

APPENDIX II

DE NOVO DERIVED PEPTIDE SEQUENCES FROM *HALORHODOSPIRA HALOPHILA* PROTEINS DISCUSSED IN PART 3.3.1

Appendix IIa. *De novo* derived peptide sequences for the identification of proteins from Halorhodospira halophila *(*yellow light) Figure 3.19a

Spot[a]	Protein[b]	Identification FASTS[c]	Sequence	E-score[c]	MS BLAST score[d]	MS- Hom. score[e]
1,2	71909086	porin, Gram-negative type	AFSV GDFGTIR AIGYDHAITTR DSYVGIEGDFGTIR TTAYAVYAHMDNDE	3.50e-05	135	124
3	37527547	S-adenosylmethionine synthetase	QDV[168]SR VYVNVGR EDIDV[283]ER AAAIKDAAGIK KIIVDTYGGMGR YVAKNVVA[241]ADR	2.20e-17	202	216
4	39936346	elongation factor Tu	GQVICKPK GITIATAGAT VIAEAHG[172]AR KTTCTGVEMFR KIIDQGEAGDNIGAIIR GVIKVGEEVEIVGITDTR	2.90e-29	336	376
5	26991638	fructose-1,6-bisphosphate aldolase	IKPIAIE IKEIHQR KVNIDTDIR KYAGVPFYR	6.60e-14	115	186
6	33634721	phosphoribulokinase	NAFE FRQPER EQVTPVVIEGDA GYTAEKTVD[228]VR	5.40e-05	95	129
7	9392587	sarcosine-dimethylglycine methyltransferase	DYAVR VHKEIER YIAHTYGCR SQYD[244]AIEVAR	0.00021	-	113
8	34497819	acetoacetyl-CoA reductase	ENWDAVMR QQIVDTIPVRR SGHTVVTTY[287]DGR SQKTAIVTGGIGGIQAVCER	4.30e-10	149	159
9	74317158	triosephosphate isomerase	YVIVGHSER VAGKFVAAR TATPEQAEAVH	7.80e-12	156	121
10	34498798	adenylate kinase	VYHVPF AAVKAGTP VAQADCADGFIFDGFPR	1.10e-15	177	191
11	53804425	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	AQINP KAVIISFR AKVGINEA EAVEEAIRR AAQIQKVIDEAFEQR	1.40e-08	167	144
12	47574096	pentose-5-phosphate-3-epimerase	KEIEAAHR RIIDEIN[225]R IEVDGGIKAENIR	0.00046	105	98

13	67941974	superoxide dismutase	VSNI SAFGSF AYQIAGS VWEHAYY QEFTQAAIGR FGSGWAWIIK IDGN[138]WNHDI EGTPIIGIDVWEHA	5.70e-17	313	298
14	68304953	DsrC	EYYDEYQIAPAVR	1.60e-07	99	82
15	78700374	nucleoside diphosphate kinase	FFFKADEIFTR EIMGATNPKEAAAGTIR TISIIKPDAVAQNAIGEIIAR	4.00e-29	279	259
16	132132	ribulose bisphosphate carboxylase small chain	IIGYDN VYDSKISDNPSR RFETFSYI[226]YGIDADVR	1.10e-06	107	113
17	53805138	pterin-4-alpha-carbinolamine dehydratase	NFIDAISFVNR EETHKIEGIHENDFIIAA VTEVAEAEDHH[213]IIIGYG SASGVKTCV...GGVAGVDAR	4.00e-13	188	132
18	69951812	cold-shock protein, DNA-binding	AIDEGQAVEF ESGVDVFVHFR	2.70e-08	112	115

Appendix IIb. *De novo* derived peptide sequences for the identification of proteins from *Halorhodospira halophila (* green/blue light) Figure 3.19b

Spot[a]	Protein[b]	Identification FASTS[c]	Sequence	E-score[c]	MS-BLAST score[d]	MS- Hom. score[e]
1	77166263	chaperone protein dnaK (Hsp70)	AKIESIVED RFDEDVVQR QSVTNPQNT FDITID[281]PR	1.60e-11	186	155
2	54294307	30S ribosomal protein S1	RAVVEEEGG AFIGPSIVDIRPVR VKGGFTVDIG[250]R	1.60e-13	152	181
3	53762519	chaperonin GroEL (HSP60 family)	GSAFEEKGR VEDAIHATR VAAVKAPGF[172]R AAVEEGIVPGGGTAIIR QIVIDA[186]DASVVPDKVR AIAAIEGDSENEEQTQGIAIVR	8.60e-26	383	299
4	71899446	ATP synthase F1, alpha subunit	RIGDAFDAAAEAR QIAAFAQFASDIDQTTRQEIER	1.00e-10	152	127
5	71550918	ribulose-bisphosphate carboxylase	VAIEA VITKIIR AIEIEDVR IGNSAKNYGR	1.60e-05	99	160
6	37527547	S-adenosylmethionine synthetase (methionine adenosyltransferase) (AdoMet synthetase)	EIFDIR TGGPI MIGIEQPI AAAIKDAAGIK KIIVDTYGGMGR YVAKNVVAAGIADR	1.00e-22	194	257

7	56461311	fructose/tagatose bisphosphate aldolase	KYA YEAF IKEIH IKPIAIET KVNIDTDIR	4.40e-11	109	156
8	78702072	phosphoribulokinase	FRQPER EQVTPVVIEGDAFHR NISHFGPEANVF[186]IEN	1.40e-16	158	191
9	33152351	NADH- dependent enoyl-ACP reductase	IIVTG YMAADLG QGIAITITYINDK TSHEISAYSFVAIAR	1.20e-12	161	190
10	2497482	adenylate kinase (ATP-AMP transphosphorylase)	IN[172]VEIQV VYHVTHNPPR AAVKA[1580]PL VAQAFAQD[204]IFD[204]PR	3.10e-09	174	195
11	52006362	dissimilatory sulfite reductase	EYYPFYQIAPAVR	0.00097	73	70
12	77866837	nucleoside diphosphate kinase	QDMGATN TISIIK[212]AVAQNAIGEIPSR	4.70e-10	101	132
13	1402737	major cold-shock protein	EGGEDVFVHFR	9.70e-05	77	71

[a] spot number according to the position on the PAGE-gel (Figure 3.19)
[b] NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/)
[c] based on the lowest E-value in FASTS database searches
[e] MS-Blast score for searches against a nr database at http://dove.embl-Heidelberg.de/Blast2/msblast.html
[f] MS-Homology score against the NCBI on-redundant protein database (Protein Prospector 4.0.5)

APPENDIX III

DE NOVO DERIVED PEPTIDE SEQUENCES FROM MUSA SP. PROTEINS DISCUSSED IN PART 3.3.2

Appendix IIIa *De novo* derived peptide sequences for the identification of proteins isolated from *Musa* variety ITC 0084

Spot[a]	Protein[b]	Identification FASTS[c]	Sequence	E-score[c]	MS-BLAST score[d]	MS- Hom. Score[e]
1, 2	7330642	HSP68	QPGVD DVATVD THGDDAQR QAVTNPTN V[I|L]V[I|L]VE SKFET[I|L]VNH MVGVSK[I|L]VR [I|L]AG[I|L]DVQR AV[I|L]TV[168]YFND	3.2e-25	277	223
3	52353541	put. ketol-acid reductoisom.	GVAYMV VT[I|L]AM KGSSSFDEAR MWKVGEQVR EKVT[I|L]AGHDE	4.9e-18	183	226
4	12546	chaperonin 60	V[I|L]E[I|L]A S[I|L]EFKDR [I|L]TV[I|L]EQ G[I|L]TMAVDSVVTN GY[I|L]SPYFITDQKNQR	4.2e-19	223	244
5	31432537	mito. chaperonin 60, HSP60-1	[I|L]AKN DDVWR S[I|L]EFKDR G[I|L]SMAVDSVVT	6.8e-08	121	124
6	4388534	F1-ATP synthase	VVD[I|L][I|L]A Q[I|L]SE[I|L]G[I|L]Y M[I|L]SPHV[I|L]GED VG[I|L]TG[I|L]TVAEHFR V[I|L]NTGS[210]TDPVGR FTQANSEVSA[I|L][I|L]GR [I|L]V[I|L]EVAEH[I|L]GEN D[I|L]NV[I|L][187]PIDQGDGE[I|L]D VDHF[I|L]F	1.2e-41	579	427
7	3746942	actin 1	GY[I|L]FTTTAER [I|L]KVVAP[226]R AVFPS[I|L]VGRPR VAN[I|L]EH[211][I|L][I|L]TEAP[I|L]N	1.7e-23	215	270
8	52076544	put. Cytosolic phosphogly kin.	VTDDTR FNEQQ[I|L] [I|L]ASQMD VVA[I|L]AV YS[I|L]K[196]VPR [I|L][I|L][I|L]ASH[I|L]GR	5.3e-08	141	162
9	57014097	pectinesterase 3 precursor	[I|L]TAA [I|L]ENTAGPSKHQ SAT[I|L]AVVGEGF[I|L]AR	2.7e-12	153	163
10			V[I|L]DVY YGPASGPD	-	-	-
11 & 12	10716961	polyphenol oxidase	DSVFFCHH YY[I|L]HFYER	2.1e-06	111	112
11			MGED[I|L]YR	-	-	-

			MSDSGENPTK[I|L]R VPEFGTTAQA[I|L]DASDKGR			
12	18479040	26S proteasome reg. subu. IV	YDAHSGGER [I|L]FQ[I|L]HTSR	1.8e-09	124	117
12	12802327	mito. proc. peptid beta subu.	YASPHPA[I|L]ADHT[186][I|L]AAPET R	8.9e-10	108	113
13	33113259	enolase	[I|L]AKYN GNPTVEVD[I|L]HCDDGT [I|L]EEE[I|L]GAAAV[171]DP	3.8e-13	210	181
14	6601496	S-adeno.-homocys. hydrolase	HS[I|L] AEFG [I|L]V[I|L]KN[I|L] SKFDN[I|L]YGCR [I|L]VGVSEETTT[173]KR WCSCN[I|L]FSTQDHAAAA	1.4e-32	312	265
15	114411	ATP synthase alfa chain	VVSV[214]QAR VVDA[I|L]GV EAFP[172]VFY[I|L] QMS[I|L][I|L][I|L]R AAE[I|L]TT[I|L][I|L] MTNFYTHFQVSE[I|L]GR	1.6e-27	304	282
16			Y[I|L] F[I|L]VF [I|L]G[I|L]S[I|L]FASVR	-	-	-
17			S[I|L][I|L]GC SYDYM[I|L][I|L]S[I|L]R [I|L]CDAADGKS[I|L][I|L]GCDEAR	-	-	-
18	37020723	ascorbate peroxidase	GEPDV G[I|L][I|L]AEKNCAP[I|L] FPAE[I|L]AHGADDG[I|L]N[I|L] DVFGHMG[I|L]SDQD[I|L]VA[I|L]SGG H	5.7e-26	266	282
19						
20			YKA[I|L]EGGHGIR VAGCCT[244]VTVNVSVHR	-	-	-
21	37783265	ascorbate peroxidase	[I|L]AGEHYYQ [I|L]EAESAHGANDG[I|L]D[I|L]AVR	2.0e-07	92	97
22			QYQ[I|L]P[I|L]QR	-	-	-
23	26453278	put. succ. dehydr. Flavoprot.	TEDGK[I|L]YQR [I|L][I|L]GE[I|L]EDY SSYT[I|L]VDH[264]DAVVVGA	1.5e-16	172	191
24			FRGEMSR SWCAVYSAR	-	-	-
25			V[I|L]YA GV[I|L]YR GGA[I|L]N[I|L][269]R	-	-	-
26			V[I|L]G [I|L]GDSDSSH VFDA[I|L][214]GR	-	-	-
27	2369714	elongation factor 2’	GFVQFCYE	9.8e-25	271	249

			VKFT[200]E[I|L]R AMKFSVSPVVR GGGQ[I|L][I|L]PTAR			
28			[I|L]DGV[I|L]H[I|L][I|L]PR	-	-	-
29	50909007	putative elongation factor 2	V[I|L]K[I|L] AMKFSVSP VKFT[201]E[I|L]R	5.4e-06	111	108
30						
31	27650423	ascorbate peroxidase	[I|L]PDA GF[I|L]AEKDCA [I|L]EAE[I|L]AHGADDG[I|L]D[I|L]	7.0e-11	139	158
31	47607439	mit. ATP synthase precursor	TY[I|L][I|L]T[I|L]QQ[I|L]R [I|L]KYT[I|L]EQH[229]G[I|L]	4.4e-04	104	107
32	4336905	ran-related GTP binding prot.	DGYYDH [I|L]TYKNV FYCWDTA[185]EKFGG[I|L]R	6.9e-17	184	199
33	7435012	14-3-3 protein tft6	Y[I|L]AEFKTGAER [I|L][I|L]SS[I|L]EQKEESR [I|L][I|L][I|L]SVAYKNV[I|L]GAR	5.3e-26	258	216
34	1658313	osr40g2	WYKDMR VYT[225]D[I|L] INFDAYHGDKDH[213]R [I|L][I|L]TKAGPDYS[I|L][200]R	3.0e-10	145	205
35	2286153	cyto. malate dehydrog.	V[I|L]VT E[I|L]VSDDDW[I|L]KGEF[I|L]TTVQQ R	6.3e-14	137	113
36	1527223	glutamine synthetase	HETAD[I|L]NTF[I|L]WGVANR	2.2e-12	126	111
37	50932771	put. malate dehydrog.	AKTF ANEDSDVV[I|L][I|L] K[I|L]FGVTT[I|L]DVVR	6.2e-11	132	135
38			[I|L]VT[I|L] FGHAA[I|L]V[I|L]	-	-	-
39	6136112	UTP-gluc-1-phos. Uridyltransf.	FFDHA[I|L]G[I|L]NVPR	1.4e-06	94	86
40	33113259	enolase	[I|L]EEE[I|L]GAAAV [I|L]AKYNQ[I|L][I|L]R	4.8e-09	130	111
41	4206124	t-comp. prot1 ε-subunit	S[I|L]HD[184]CVAR M[I|L]Y[I|L]EHCANSR	5.5e-13	151	146
42 & 43	56554972	heat shock protein 70	MVDKMD ARFEE[I|L]NMD[I|L]HR QATKDAGV[I|L]AG[I|L]NV[I|L]R	2.7e-16	186	128
44			[I|L]T[I|L]Y[I|L]KR	-	-	-
45			GQ[I|L]S[I|L]E FAEQ[I|L]KDR	-	-	-

Appendix IIIb. ***De novo*** derived peptide sequences for the identification of proteins isolated from *Musa* variety ITC 0643

Spot[a]	Protein[b]	Identification FASTS[c]	Sequence	E-score[c]	MS-BLAST score[d]	MS- Hom. Score[e]
1						
2			EYAWRES[I|L]	-	-	-
3	41818408	class III acidic chitinase	FGMGQT NSGY…KSHV [I|L]SSE[I|L]QSC	1.3e-09	163	148
4						
5	39939493	ascorbate peroxidase	A[I|L][I|L]TD [I|L]PDGREYD G[I|L][I|L]AEKNCA FPAE[I|L]AHGADD[170]S[I|L] DVFGHMG[I|L]SDED[I|L]VA[I|L]SN	7.3e-16	203	211
6	1296955	r40c1 protein	V[I|L]ASV WYKDMR HDD[I|L]SPR VYTKADPNYS DGTT[I|L]VNWE [I|L]FTKAGNDYS HA[I|L]EKNP[I|L]K [I|L]NFDAFHGDKDH	2.6e-17	226	209
7	37928995	cytosolic malate dehydrogen.	Q[I|L]VQGDD A[I|L]GQ[I|L]DTR GAA[I|L][I|L]KAR [I|L]DHNRA[I|L]G	2.3e-07	171	120
8	56202334	alpha-amylase isozyme III	GNYC[I|L]A [I|L]YD[I|L]D TDNGFD[243]R KQGGWYNF[I|L]R	3.2e-09	72	191
9			YS[I|L]K	-	-	-
10	25809056	DEAD box RNA helicase	[I|L][I|L]SSG G[I|L]YAYGFEKPS	2.9e-07	84	105
11			KAFDEG	-	-	-
12						
13			DSVFFCHH MWD[I|L]YR DGTNNPTK[I|L]R	-	-	-
14	55297085	put. ketol-acid reductoisom.	GVSFMV DS[I|L]AAADSD[I|L]VV[266]G[I|L]R	3.7e-07	80	105
15			S[I|L][I|L]YS	-	-	-

[a] spot number according to the position on the PAGE-gel (Figure 3.23)
[b] NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/)
[c] based on the lowest E-value in FASTS database searches
[e] MS-Blast score for searches against a nr database at http://dove.embl-Heidelberg.de/Blast2/msblast.html
[f] MS-Homology score against the NCBI on-redundant protein database (Protein Prospector 4.0.5)

APPENDIX IV

DE NOVO DERIVED PEPTIDE SEQUENCES FROM SHEWANELLA ONEIDENSIS PROTEINS DISCUSSED IN PART 3.3.3.1

Appendix IVa. SPITC

Spot[a]	Protein[b]	Identification[c]	Sequence	Scores FASTS[c]	MS BLAST[d]	MS- Hom[e]
1	-					
2	24376191	hypothetical protein SO4719	ESKTAEEAFAK FVADGFGVEPR ANELQGYTLSDR SGLPFLSR	7.3e-28	298	270
	24371835	ribosomal protein S3 S.o.	LAGTPAQLNLAELR GLKVEVSGR	5.9e-14	163	128
	78692243	alpha keto acid dehydrogenase complex, E1 component, beta subunit	ELVKR VGHFG SGNEFNVGSLVFR	2.8e-10	135	123
3	24375384	outer membrane porin, putative	KTNAGTVLVGR SAD[156]WYYSPK LSVELFNVQGAYR KSTMVYGQYSMYR GNVDVFGNTNADLDR VAGN[270]DLQDDNVFS[255]R	4.5e-66	554	518
	69953123	ketose-bisphosphate aldolase, class-II:	KVNLDTDLR AAEATDS[196]LVQASAQR	1.9e-10	148	132
	24374021	alcohol dehydrogenase, iron-containing	GSFVQLDDVLAAQR	4.0e-07	97	88
4	24375384	outer membrane porin, putative	VYGQYSMYR TNAGTVLVGR SADGVWYYSPK S[331]YGQYSMYR LSVQNFNVGADYR DLQDDNVFS[255]R	3.8e-45	401	446
	78365797	ketose-bisphosphate aldolase, class-II:	LPN[238]LVMHGSSSVPQE A[244]QAAEATDS[196]LVQASAGAR	1.0e-26	227	239
5	82498382	translation elongation factor Tu: Small GTP-binding protein domain	KLLDEGR NTSHLEYDT[184]R V[172]EVELVGLR TTDVTGTLELPEGVEMVMPGDNLK	7.3e-42	359	334
	78690797	outer membrane porin, putative	KTNAGTVLVGR S[186]GVWYYSPK LSVQNFNVGADYR	1.2e-25	243	231
	69953123	ketose-bisphosphate aldolase	EATDS[196]LVQASAGAR	4.9e-07	93	93
6	24371815	translation elongation factor Tu	EHLLLSR KLLDEGR TGVEMFR VGDEVELVGLR DFSQLDNAELER MPLEDVFSLSGR GLTLNTSHLEYDTPSR L[242]LAAALDSYLPEPER TTDVTGTLELPEGVEMVMPGDNLK	2e-95	797	742
	77816497	citrate (Si)-synthase	LMGFGHR DFVDLDKR LFLLHADHEQNASTSSLR	1.6e-22	227	208
	24374618	long-chain fatty acid transport protein	AVLADNATVLSR	5.9e-05	79	59
7	68545906	IMP dehydrogenase	DDAADLKVPEGIEGR	4.0e-18	155	174

			LNLPLVSAAMDTVTEAR			
	82744041	trigger factor	KQHATYA F[242]VANMPELPAELFTEQAAR	2.3e-18	131	182
	24372021	dihydrolipoamide dehydrogenase	TVLVER YDAVLVALGR G[260]NVDKQLR	4.8e-13	125	154
	82495877	phosphoenolpyruvate carboxykinase (ATP)	A[243]PLEHLQAR V[330]YAWHQLFAR	2.9e-07	122	92
	82744041	GTPases - translation elongation factors	EL[226]GVEMVMPGD[227]K	8.6e-07	92	109
8	24372295	chaperonin GroEL	VVFGDDAR VEDALHATR ELLPLLEGLAK AAVEEGVVPGGGVALLR V[185]EDQKHGVVL[184]R DNTTLLDGDGEQAQLEAR	1.5e-56	496	461
9	24372557	fumarate reductase flavoprotein subunit precursor	AAAVSAR FMNELTTR NAAETKPQAK L[229]ALSDLVTYGR PGATGDGLDVALQA LQAHPTYSPAGGVML	1.8e-51	452	413
10	78369113	ribosomal protein S1	VDVR[196]R YPENTKLTGR	1.7e-06	112	114
11	78692288	phosphoenolpyruvate synthase	SREDVQLLER LEFLLNR MLGLHPK Q[347]NDAEVMELAK L[228][170]ASLGSAFYPK E[364]KLAAETSDASFAVR Q[285]LDT[244]QPALEQALR	5.1E-57	519	429
12	-					
13	-					
14	24374798	phosphoribosylformylglycinamidine synthase	LGAVLQVSR YVESDTLTAEQQR T[281]LLDFGASAR	2.0e-20	220	212
15	24374136	hypothetical protein SO2593	SQVWQK ELPLLNEKVQR G[200]QLKEQLR A[242]FLPSEEELTER A[341]LTDNLLDGELVH	1.4e-31	348	340
16	24376191	hypothetical protein SO4719	GTFVAYK SGLPFLSR ESKTAEEAFAK FVADGFGVEPR T[360]ANELQGYTLSDR	9.1e-40	357	324
	78692093	ribosomal protein S3	VRQPR VPLHTLR KGEDVEVLR L[273]GLKVEVSGR L[229]PAQLNLAELR	1.1e-25	230	173
17	82744520	ribosomal protein L2	VGNAEHMLR N[196]DH[234]G SAGAYVQVVAR	3.8e-14	161	186
	82744371	translation elongation factor Tu:Small GTP-binding protein domain	VGD[228]ELV[170]R [184]AALDSYLPE[226]R	4.1e-08	119	152
	24373887	alpha keto acid	S[300]FNVGSLVFR	0.0012	71	84

		dehydrogenase complex, E1 component,beta subunit				
18	78688396	translation elongation factor Tu:Small GTP-binding protein domain	VGDEVELVGLR AAALDSYL[226][226]R	3.6e-11	135	144
	69950403	ribose-phosphate pyrophosphokinase	F[201][186]ISVQLNENVR	0.0008	82	82
19	68544347	malate dehydrogenase, NAD-dependent, eukaryotes and gamma proteobacteria	F[188]SLVR SDLFNLNAGLVR L[303]TTLDVLR VL[230]GEVSAFEADAR	2.4e-25	252	216
	24371827	translation elongation factor Tu	V[172]EVELVGLR AALDSYLPE[226]R	5.1e-10	131	139
	78687894	inorganic diphosphatase	L[186]LS[226]TAFLLER	0.033	64	65
20	24375384	outer membrane porin, putative	E[241]FYGR SADGVWYYSPK S[202][158]DLSVQNFNVGADYR	1.5e-24	238	242
	24374021	alcohol dehydrogenase, iron-containing	G[234]VQLDDVLAAQR	0.0001	82	88
21	-					
22	-					
23	-					
24	24371827	translation elongation factor Tu	KLLDEGR VGDEVELVGLR G[214]LDTSHLEYDTSPR L[243]LAAALDSYLPELPR DLDK[245]LM[210]EDVFSLSGR	8.0e-40	384	356
	77816497	citrate (Si)-synthase	D[246]DLDKR L[260]LHADHEQDAST[190]PR	1.1e-08	125	136
25	78688058	dihydrolipoamide dehydrogenase	F[283]AASGR YDAVLVALGR GFLNVDKQLR A[186]LGLETVLVER	4.5e-24	252	256
	63079040	elongation factor-Tu-2	EHLLLSR K[226]DEGR V[172][288]ELVGLR EL[382]EMVMPGDNLK	2.2e-19	246	193
	78506850	phosphoenolpyruvate carboxykinase (ATP)	A[234][210]EHLAQR V[330]YAWHQLFAR	1.3e-09	131	135
	78069867	citrate synthase	LMGFGHR L[226]FLAR	0.00052	82	93
26	24372295	chaperonin GroEL	VEDALHATR ELLPLLEGLAK EDQKHGVVLALR A[170]EEGVVPGGGVALLR DN[315]LDGDGEQAQLEAR	6.7e-43	415	399
	24373894	glyceraldehyde-3-phosphate dehydrogenase	D[198]LYGFGR ALDHADDLAPR	3.6e-10	129	132
	24371827	translation elongation factor Tu	HLEYDTPSR VA[330]ELVGLR	3.4e-07	113	104
27	24372557	fumarate reductase flavoprotein subunit precursor	AAAVSAR FMNELTTR L[228]ALSDLVTYGR A[215]HPGATGDGLDVALQA[300]R	2.0e-29	288	301
	24371815	translation elongation factor Tu	ELVGLR NTSHL[292]DTSPR	0.00011	93	91

28	24373579	heat shock protein 90	Y[297]TNDALYEGDGELR F[228]GLLDSNDLPLNVSR	3.9e-23	205	211
	78369113	ribosomal protein S1	YPENTKLTGR AFL[241]LVDVRPVR	2.7e-13	152	160
29	78506612	phosphoenolpyruvate synthase	LEFLLNR MLGLHPK ASLGSAFYPK S[285]DVQLLER L[340][156]GLAR D[257]LSHLFDER QF[200]ND[200]VMELAK Q[398]DT[244]QPALEQALR VHQ[220]EHHGVALSAG[227]R	3.7e-56	509	655
	78367230	translation elongation factor G:Small GTP-binding protein domain	FGALTFVR L[260]VNKLDR T[154]HVDFTVEVYR T[319]ELLVGEPQVAYR	1.7e-27	285	282
30	24374136	hypothetical protein SO2593	LN[186]LVH VE[156]HLR G[372]SVKR YLDLLNATLR LLLASGLTGR EVFGLAELTK G[200]QLKEQLR FL[170]YASNVYNR A[242]FLPSEEELTER QLDATFSQAYLEETFGR	5.8e-65	641	618

Appendix IVb. 2-sulfobenzoic acid cyclic anhydride

Spot[a]	Protein[b]	Identification[c]	Sequence	Scores FASTS[c]	MS-Blast[d]	MS- Hom[e]
3	82497768	outer membrane porin, putative	FSKSTMVY	0.033	63	49
4	82497768	outer membrane porin, putative	FSKSTMVY SADGIWYYSPK	4.8e-13	145	131
6	78688396	translation elongation factor Tu:Small GTP-binding protein domain	MKVTLLC[210][202]DE[170]R DLDK[244]LM[210]EDVF[200]SGR	9.3e-15	162	230
7	78506850	phosphoenolpyruvate carboxykinase (ATP)	VTTQYAWH	0.37	67	60
8	68544820	glyceraldehyde 3-phosphate dehydrogenase	LAV LRESF FEMAE	0.71	78	nr
11	27361244	translation elongation factor G	VKADV[210]SEMF	0.0091	69	70
15	78685062	NAD-glutamate dehydrogenase	TLGKFME[241]NLR	0.00033	74	76
30	24374136	hypothetical protein SO2593	KEALDHA AAFREELDL	0.048	101	89

[a] spot number according to the position on the PAGE-gel (Figure 3.28)
[b] NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/)
[c] based on the lowest E-value in FASTS database searches
[e] MS-Blast score for searches against a nr database at http://dove.embl-Heidelberg.de/Blast2/msblast.html
[f] MS-Homology score against the NCBI on-redundant protein database (Protein Prospector 4.0.5)

APPENDIX V

DE NOVO DERIVED PEPTIDE SEQUENCES FROM MUSA BALBISIANI PROTEINS DISCUSSED IN PART 3.3.3.2

Spot[a]	Protein[b]	Identification FASTS	Sequence	FASTS[c]	MS BLAST score[d]	MS-Hom. Score[e]
234	11066033	cytosolic aconitase	LLLESALR LVELPFKPAR	1.8e-07		101
266	77556324	putative heat-shock protein	GFAHPER YEYQAEVSR DVTTEEYNEFFR DLYYLAADSLVSAR	1.7e-18	228	198
397	24637539	heat shock protein 60	ELGELLAK V[242]LSLKR V[271]DGVTVAK NVVLEQSFGAPK	8.1e-17	140	177
453	37531422	putative enolase (2-phospho-D-glycerate hydroylase)	L[258]ELGAAAVYAGAK GLPTVEVDLHCDDGTFAR AAVPSGASTGVYEALELR	1.4e-32	293	288
	2645893	F1 ATPase a-subunit	LLSQYER VFYLHSR A[200]LTTLLESR	3.2e-09	156	139
468	22273	enolase	ENQLLR YEALELR AGWGVMASHR D[201]DYLGKGVLK LEEELGDAAVYAGVKFR A[170]PSGASTGVYEALELR	2.2e-39	381	434
571	33339126	actin	AVFPSLVGR LKVVAPPER GYSFTTSAER LA[262]ALDYEQELEAAR VANNQHPLLLTEAPLNPK	4.1e-40	395	365
575	1498395	actin	AVFPSLVGR GYSFTTTAER LKVVAP[226]R DFSVGDEAQSKR VA[226]EHPLLLTEAPLNPK	6.6e-36	320	351
583	20513166	3-isopropylmalate dehydrogenase	LEAAVLNTLNR ELTGGLYFGKPR	9.7e-08	103	114
	3738259	cytosolic phosphoglycerate kinase 1	YLLEHGARVLLCSHLGR	1.6e-06	107	108
595	50251688	putative aspartate transaminase	YYHPEVR DDNGKPVVLECVR	9.5e-12	144	140
	1297361	alcohol dehydrogenase 1	THPVNFLNER	7.9e-05	77	70
606	60101357	glutamine synthetase	NDGGYEVLK DLVDAHYK ND[277]EVLKSALEK	3.6e-14	157	134
614	50940085	putative r40c1 protein	KDMR HHHHHHR DGTTVVLWEWLKGDNQR	3.2e-11	169	188

615	1620972	L-lactate dehydrogenase	LHPVSLLAK K[170]VDSAYEVLR EVFLSLPAQLGR	1.1e-14	187	169
620	1658313	osr40g2	ADPNYSLSLR L[261]DAFHGDKDHGGVR	5.8e-10	133	129
630	10798652	malate dehydrogenase	ALGQLSER TPAGEKPVR L[213]QVSDVK K[287]ALSAASSACDHLR	1.6e-26	259	261
638	30060226	1-aminocyclopropane-1-carboxylic acid oxidase	FYDSELAK LLYPGGYR TFAPPFVGTK FMLYLHYYFGTK VPVLDLAEFESEER	1.1e-15	197	238
674	76559896	TPA: isoflavone reductase-like protein 6	FLVFASAR YTTVDEFLNR LGNPTFALVR VYVPEEEVLK YLPSEFGNDVDR ALFLNEDDLGTYTLK	2.9e-30	308	320
682	15705988	endochitinase	P[247]HDVLTGR WTPSNADQAAGR G[337]NSFNYNYGPAGR	2.6e-23	226	213
709	7739434	14-3-3-like protein	VEFMEK LVSSLEQKEESR NLLSVAYKNVLGAR	4.2e-22	221	179
711	41818408	class III acidic chitinase	SHVCPAR SFLDGSAAR LSGYSSQVK LSSELQSCQER	1.2e-20	237	215
712	1668706	atran2	KPFLYLAR NVPTWHR NLQYYELSAK QQTVDYPSFK	9.1e-26	269	241
739	2586151	ripening-associated protein	LPDVGKGSDHLR FP[200]LAHGANNGLSLAVR	8.6e-17	164	179
744	76573375	triosephosphate isomerase-like protein	VATPAQAQEVHYELR V[184]CV[186]TLEQR	6.9e-15	163	166
746	76573375	triosephosphate isomerase-like protein	VLACV[186]TLEQR VATPAQAQEVHYELR	7.9e-17	174	163
769	47575681	abscisic stress ripening protein-like protein	KHEAK KDPDHAHK	8.6e-06	62	96
772	601871	manganese-superoxide dismutase	VATVSLPR K[214][228]TTANQDPLVTK	8.2e-07	102	121
784	3492854	mitochondrial small heat shock protein	KEVFQV EDEHALHLR EDGNVDVDLER ADSFPSLQEVFDPFNPTR	3.0e-06	66	nr
788	34334012	cytosolic glutathione peroxidase	ANSAGTLHDFTVK F[226]DKDGHVVDR	9.1e-24	225	223

			YAPTTSPLSLEKDLK			
813	47026989	nucleoside diphosphate kinase	GLVGELLNR NVLHGSDSLESAR NVLHGSDSLEGASK KLLGATNPADSAPGTLR	8.4e-25	234	254

[a] spot number according to the position on the PAGE-gel (Figure 3.32)
[b] NCBI Entrez entries (http://www.ncbi.nih.gov/Entrez/)
[c] based on the lowest E-value in FASTS database searches
[e] MS-Blast score for searches against a nr database at http://dove.embl-Heidelberg.de/Blast2/msblast.html
[f] MS-Homology score against the NCBI on-redundant protein database (Protein Prospector 4.0.5)

Not every **end** is the goal. The **end** of a melody is not its goal, and yet if a melody has not reached its **end**, it has not reached its goal. A parable.

Nietzsche, Friedrich (1844-1900)