DEVELOPMENT OF TOOLS FOR PROTEOME ANALYSIS

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH
for the degree of
DOCTOR OF NATURAL SCIENCES

presented by

MARTIN CHRISTIAN MÜNCHBACH

Dipl. Natw. ETH
born August 29, 1970
Federal Republic of Germany

accepted on the recommendation of
Prof. Dr. Hauke Hennecke, examiner
Prof. Dr. Ernesto Carafoli, co-examiner
Dr. Peter James, co-examiner

Zürich - 1999
# Table of Contents

## Introduction

1. **Chapter:** Genomics
   - 1.1 Comparative gene expression analysis - Functional Genomics
     - 1.1.1 Hybridisation based techniques
     - 1.1.2 PCR-based techniques
     - 1.1.3 Sequence-based techniques
   - 1.2 Proteomics
     - 1.2.1 Technology for Proteomics
       - 1.2.1.1 Separating and visualising proteins
       - 1.2.1.2 Detecting proteins
       - 1.2.1.3 Identifying proteins
       - 1.2.1.4 Processing peptide data
     - 1.2.2 Application of the technology
       - 1.2.2.1 Functional Proteomics
       - 1.2.2.2 Cell-map Proteomics
   - 1.3 Economics
     - 1.3.1 More effective drugs
       - 1.3.1.1 Pharmacogenomics
       - 1.3.1.2 Pharmacogenomic technologies
       - 1.3.1.3 Pharmacogenomic developments
     - 1.3.2 Increasing research effectiveness
   - 1.4 References

## Separating and Visualising Proteins

2. **Chapter:** Separating Proteins
   - 2.1 Sample preparation
   - 2.1.2 The first dimension
   - 2.1.3 The second dimension

1-1

1-3

1-5

1-6

1-9

1-10

1-11

1-12

1-13

1-14

1-16

1-17

1-18

1-19

1-20

1-21

1-22

1-23

2-1

2-2

2-2

2-3

2-3
TABLE OF CONTENTS

2.1.4 Carrying out two dimensional gel electrophoresis 2-4
  2.1.4.1 Materials 2-4
  2.1.4.2 Buffers 2-5
  2.1.4.3 Cell extract preparation for B. japonicum 2-6
  2.1.4.4 The 1st Dimension
    Immobiline rehydration and loading of the sample 2-6
  2.1.4.5 The 2nd dimension
    slab gel casting and transfer on second dimension 2-7

2.2 Visualising Proteins 2-9

2.3 Application of the Techniques 2-11
  2.3.1 The biological problem 2-11

2.4 Discussion 2-14

2.5 Reference 2-16

3. CHAPTER: DETECTING PROTEINS 3-1

3.1 Mass Spectrometry 3-3
  3.1.1 Sample Ionisation
    3.1.1.1 Electrospray ionisation 3-3
    3.1.1.2 Matrix assisted laser desorption and ionisation 3-4
  3.1.2 Mass Analysers
    3.1.2.1 Time-of-flight mass spectrometer 3-4
    3.1.2.2 Quadrupole mass spectrometer 3-5
    3.1.2.3 Ion trap mass spectrometer 3-5
  3.1.3 Detectors 3-5
  3.1.4 MALDI-TOF-MS 3-6

3.2 Immunodetection 3-7
  3.2.1 Solid phase peptide synthesis - the Merrifield technique 3-8
  3.2.2 Fmoc and Boc chemistry in SPPS 3-8
  3.2.3 Protecting groups and cleavage strategies in Fmoc chemistry 3-9
  3.2.4 Coupling methods 3-10
  3.2.5 Resin tests 3-11
  3.2.6 Fmoc resin cleavage and deprotection 3-13
  3.2.7 Post-cleavage work-up 3-13
  3.2.8 Antibody generation 3-13

3.3 Application of the Techniques 3-15
  3.3.1 The synthesis of three peptides
    RpoH₁, RpoH₂ and RpoH₃ 3-15
  3.3.2 The generation of the three antibodies
    α-RpoH₁, α-RpoH₂ and α-RpoH₃ 3-21
  3.3.3 The application of the three antibodies
    α-RpoH₁, α-RpoH₂ and α-RpoH₃ 3-23
### TABLE OF CONTENTS

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Discussion</td>
<td>3-25</td>
</tr>
<tr>
<td>3.4.1</td>
<td>The detection problem</td>
<td>3-25</td>
</tr>
<tr>
<td>3.4.2</td>
<td>The biological problem</td>
<td>3-25</td>
</tr>
<tr>
<td>3.5</td>
<td>References</td>
<td>3-28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>CHAPTEER:</td>
<td>IDENTIFYING PROTEINS</td>
</tr>
<tr>
<td>4.1</td>
<td>Chemical Identification</td>
<td>4-2</td>
</tr>
<tr>
<td>4.1.1</td>
<td>N-terminal</td>
<td>4-2</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Ladder sequencing</td>
<td>4-3</td>
</tr>
<tr>
<td>4.2</td>
<td>Tandem Mass Spectrometry</td>
<td>4-6</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Tandem-in-space and tandem-in-time mass spectrometry</td>
<td>4-6</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Low energy peptide fragmentation</td>
<td>4-7</td>
</tr>
<tr>
<td>4.2.2.1</td>
<td>Concentration gels</td>
<td>4-10</td>
</tr>
<tr>
<td>4.2.2.2</td>
<td>Proteolytic in-gel cleavage of peptides</td>
<td>4-11</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Derivatisation - Simplifying interpretation</td>
<td>4-11</td>
</tr>
<tr>
<td>4.2.3.1</td>
<td>Ion labelling I: The $^{16}$O/$^{18}$O method</td>
<td>4-12</td>
</tr>
<tr>
<td>4.2.3.2</td>
<td>Ion labelling II: Nicotinic acid derivatisation</td>
<td>4-13</td>
</tr>
<tr>
<td>4.3</td>
<td>Application of the Techniques</td>
<td>4-16</td>
</tr>
<tr>
<td>4.3.1</td>
<td>MS/MS Sequencing</td>
<td>4-16</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Derivatisation</td>
<td>4-19</td>
</tr>
<tr>
<td>4.3.2.1</td>
<td>Derivatisation chemistry I: The Nicotinylation</td>
<td>4-19</td>
</tr>
<tr>
<td>4.3.2.2</td>
<td>Derivatisation chemistry II: The Succinylation</td>
<td>4-26</td>
</tr>
<tr>
<td>4.3.2.3</td>
<td>Derivatisation chemistry III: The effect of Nicotinylation on MS/MS</td>
<td>4-28</td>
</tr>
<tr>
<td>4.3.2.4</td>
<td>Derivatisation chemistry IV: The $^{16}$O/$^{18}$O method</td>
<td>4-32</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion</td>
<td>4-34</td>
</tr>
<tr>
<td>4.4.1</td>
<td>The identification problem</td>
<td>4-34</td>
</tr>
<tr>
<td>4.4.1.1</td>
<td>Mass spectrometry versus automated Edman sequencing: TIME</td>
<td>4-34</td>
</tr>
<tr>
<td>4.4.1.2</td>
<td>Mass spectrometry versus automated Edman sequencing: QUALITY</td>
<td>4-35</td>
</tr>
<tr>
<td>4.4.1.3</td>
<td>Mass spectrometry versus automated Edman sequencing: QUANTITY</td>
<td>4-35</td>
</tr>
<tr>
<td>4.4.1.4</td>
<td>Mass spectrometry versus automated Edman sequencing: COST</td>
<td>4-36</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Outlook</td>
<td>4-37</td>
</tr>
<tr>
<td>4.5</td>
<td>References</td>
<td>4-38</td>
</tr>
</tbody>
</table>
# Table of Contents

5. **Chapter:** **Processing Peptide Data**

5.1 Peptide Mass Fingerprinting

5.2 Sequence Tagging

5.2.1 Peptide Fragment Fingerprinting

5.2.2 Peptide Sequence Tags

5.3 Application of the Techniques

5.3.1 Peptide Mass Fingerprinting

5.3.2 Mining Genomes

5.4 Outlook

5.4.1 Cell fingerprinting

5.4.2 Sample preparation and processing

5.4.3 Data extraction from HPLC-MS/MS files

5.4.4 Creation of a cell fingerprint database with appropriate search tools

5.4.5 Potential applications

5.5 References

6. **Chapter:** **Functional Proteomics**

6.1 Multiple Small Heat-Shock Proteins in Rhizobia

6.2 Proteome Analysis of Heat-Shock Protein Expression in *B. japonicum*

**Appendix**

- Curriculum Vitae
- List of Publications
- Acknowledgement
Summary

The purpose of the work presented in this thesis was to develop new tools for Proteome analysis. The Proteome is defined as the total set of proteins expressed in a given cell at a given time under given conditions. Proteomics, the study of the function of the Proteome, has to be able to cope with a far higher degree of complexity and dynamics, than its DNA and RNA complements, Genomics and Transcriptomics, respectively. Whereas the degree of complexity at the genomic level can be increased a little by gene rearrangements such as somatic recombination, that of Transcriptomics is increased far more by processes such as intron splicing and RNA editing. Proteomics reaches a far higher level of complexity. The four-letter nucleotide code of DNA and RNA is replaced by the twenty amino acid alphabet which is complicated by the addition of one or more of the >300 types of post-translational modification.

In order to provide a conceptual framework, within which one can view the emerging field of Proteomics, this thesis describes the work carried out on the various techniques. The different techniques are presented in the individual chapters in much the same order, as one would approach a typical Proteome experiment. Firstly, the protein complement has to be separated and visualised. This is usually done by two-dimensional gel electrophoresis in combination with a variety of staining methods (separating and visualising proteins, chapter 2). As an application, a subtractive gel-analysis of a HrcA mutant is shown. The analysis confirmed the role of HrcA as negative control element in the regulatory heat shock network of Bradyrhizobium japonicum. In a next step, the proteins are characterised, either by indirect (detecting proteins, chapter 3) or direct methods (identifying proteins, chapter 4). The indirect detection approach is illustrated by in vivo stability experiments of σ32- type heat shock factors carried out by immunodetection. A detailed description of the preparation of antibodies against RpoH1, RpoH2 and RpoH3 is given, starting with the peptide synthesis of the haptens. The direct identification of proteins and peptides is demonstrated by novel approaches for sequencing by mass spectrometry. Ion labelling by nicotinic acid derivatisation allows the selective enhancement of the b-ion series in a MS/MS spectrum. The application of the ion labelling technique is demonstrated by various examples. The large amount of data accumulated during the steps described in the previous chapters must then be processed, filtered and finally interpreted to help reveal how a Proteome is acting during a particular event (processing peptide data, chapter 5). As an example for the automated detection by SEQUEST, the analysis of Grp94 is shown. In addition, an outlook is given of the steps being taken, to create a database with appropriate interrogation tools to identify cell types or bacteria (cell fingerprinting). In chapter 6, a complete example of Functional Proteomics is presented. The behaviour of B. japonicum to stress changes caused by heat shock and various other conditions is monitored. Subtractive gel-analysis revealed the induced stress proteins. A set of nineteen heat shock proteins was observed, which were characterised by the methods described in the previous chapters. In addition, the presence of multiple small heat shock proteins was demonstrated for a variety of Rhizobium and Bradyrhizobium species. Kinetic studies revealed the existence of functional groups of heat shock proteins, which differ in their heat shock response.
Zusammenfassung


Résumé

Le but du présent travail était de développer de nouvelles méthodes pour l'analyse de protéomes. Le protéome est défini comme la totalité des protéines exprimées à un moment donné dans une cellule précise. L'étude de la fonction du protéome est nommée la protéomique. La protéomique doit être capable de gérer un degré de complexité et de dynamique bien supérieur à celui des études complémentaires concernant l'ADN (génomique) et l'ARN (transcriptomique). Alors que la complexité de l'information codée au niveau génomique peut être légèrement augmentée par un réarrangement de gènes, comme dans le cas de la recombinaison somatique, celle concernant la transcriptomique est fortement amplifiée du fait de processus tels que l'épissage d'introns et l'édition d'ARN. La protéomique est encore bien plus complexe. En effet, le code à quatre lettres (nucléotides) de l'ADN et de l'ARN est traduit dans l'alphabet à vingt lettres (acides aminés) des protéines. En plus, cette traduction peut s'accompagner d'une ou plusieurs modifications posttranslationnelles, dont plus de 300 sont répertoriées actuellement.

Afin de donner un cadre conceptuel à cette nouvelle notion de la protéomique, la présente thèse décrit avant tout les différentes méthodes impliquées. La suite des chapitres correspond au déroulement typique d'une expérience de caractérisation de protéomes. Premièrement, le complément protéine est dissocié et visualisé. Ceci se fait d'habitude par une électrophorèse en deux dimensions, combinée à différentes méthodes de coloration (chapitre 2). Comme application, une analyse soustractive de gel d'une mutante HrcA est montrée. Cette analyse a confirmé le rôle du HrcA comme élément de contrôle négatif dans la régulation du processus du choc thermique chez Bradyrhizobium japonicum. Dans la seconde étape, les protéines sont caractérisées par méthodes indirectes (chapitre 3) ou directes (chapitre 4). L'approche indirecte est illustrée par des expériences de stabilité in vivo, dans lesquelles des facteurs de choc thermique σ2 sont mis en évidence par la méthode de l'immunodétection. La préparation d'anticorps contre RpoH1, RpoH2 et RpoH3 est décrite, débutant à la synthèse peptidique des haptènes. L'identification directe des protéines et peptides est montrée à l'aide d'approches nouvelles pour le séquençage par spectrométrie de masse. Le marquage sélectif d'ions par la dérivation de l'acide nucléique conduit à une intensification de la série d'ions b dans un spectre MS/MS. L'application de la technique de marquage d'ions est illustrée par différents exemples. La foule de données accumulées aux différentes étapes doit alors être traitée, filtrée et finalement interprétée pour révéler le mode d'action d'un protéome pendant un événement particulier (chapitre 5). L'analyse de Grp94 figure comme exemple d'une détection automatisée par SEQUEST. En plus, une perspective des différentes étapes nécessaires à la création d'une base de données, qui permettra l'identification de types cellulaires ou de bactéries, est décrite. Au chapitre 6, un exemple complet d'une étude de protéome fonctionnel est présentée. Cette étude décrit le comportement de Bradyrhizobium japonicum en situation de stress causé par un choc thermique ou d'autres conditions. L'analyse soustractive de gel a révélé la présence de dix-neuf protéines de choc thermique qui ont été caractérisées et identifiées grâce aux méthodes décrites aux chapitres précédents. En plus, la présence de maintes petites protéines de choc a été démontrée pour différentes variétés d'espèces de Rhizobium et Bradyrhizobium. Des études cinétiques ont révélé l'existence de groupements fonctionnels de protéines de choc thermique qui se distinguent par leur réponse au choc thermique.
1 Introduction

"You start getting into a new genome and you never know what you are going to find. It’s like Columbus discovering a New World"

CLAIREFRASER (1998)

1.1 Genomics

Genomics is the discipline that maps and tries to understand the DNA instructions that build and sustain life. Although the name is new, genomics has a rich history of prior discoveries to build on. 90AD Hippocrates observed that some illnesses and traits like eye colour and baldness, can be inherited. In 1866 Gregor Mendel, while gardening in the cloister of Brünn, postulated the basic laws of genetics. His observations of pea plants allowed him to deduce concepts such as dominant and recessive traits. In 1944 Oswald T. Avery, Colin McLeod and Maclyn McCarty showed that inherited characteristics in bacteria are determined by DNA and not protein as long believed (Avery et al., 1944). In 1952 A.D. Hershey and M. Chase established DNA as the genetic material (Hershey and Chase, 1952), while in 1953 Watson and Crick proposed the double helix structure of DNA (Watson and Crick, 1953). In 1973, Stanley Cohen and Herbert Boyer created the first recombinant DNA organism using recombinant DNA techniques pioneered a year earlier by Paul Berg (Jackson et al., 1972). 1975 Edwin Southern devised DNA hybridisation analysis (Southern, 1975). In 1977 the first complete DNA sequence of an organism, virus Φ174, was published (Sanger et al., 1977) and the first genetic engineering company, Genentech (San Francisco, CA, USA), founded. In 1978, the science writer David Rorvik published his controversial novel “In His Image: The Cloning of a Man”. The plot of the book, which Rorvik claimed was true, chronicled the fictitious story of Rorvik’s assistance to a rich man who employed a scientist to clone himself. In 1980, the United States Supreme Court ruled that a “live, human made microorganism is patentable material”, opening the doors for the creation of the modern biotechnology industry. In 1982 Gen-Bank was established and ‘Humulin’ (human insulin; Lilly/Genentech) received FDA approval. As one of the greatest achievement of modern molecular biology, Kary B. Mullis developed the polymerase chain reaction (PCR) in 1983. PCR allows the rapid synthesis of designated fragments of DNA (Saiki et al., 1985). In 1990, the National Institutes of Health (NIH)
officially began the Human Genome Project (HUGO), a massive international collaborative effort to sequence the estimated 3 billion nucleotides making up the entire human genome and locate the 50,000 to 100,000 genes. In 1991 Craig Venter developed ESTs (expressed sequence tags; see below), a way to quickly decipher expressed genes (Adams et al., 1991). In 1992 Venter and Claire Fraser founded the Institute of Genomic Research (TIGR). The new institute began a structural and comparative analysis of genomes. In 1993 the first biotech company (Incyte Pharmaceuticals) went public. Incyte provides an integrated platform of genomic technologies and software designed to aid in the understanding of the molecular basis of disease.

July 28, 1995, a team led by Robert Fleischmann, which included Venter, Fraser and Humphrey S. Smith, published, the first complete genome for a free-living organism: Haemophilus influenza (Fleischmann et al., 1995). The first eukaryotic genome Saccharomyces cerevisiae was completed April 24, 1996 (Goffeau et al., 1996) and the publication of the genome of the first Archaea Methanococcus jannaschii on August 23, 1996 confirmed the existence of the third branch of life (Bult et al., 1996). Frederick Blattner announced the complete sequence of Escherichia coli K-12, the workhorse of the molecular biologist, in 1997 (Blattner et al., 1997). An elegant piece of research was the completion of the genome sequence of the geneticists workhorse Caenorhabditis elegans (1998), the first animal and multi-cellular organism to have its genome completely sequenced. In 1999 the sequence of Drosophila melanogaster is expected to be completed, in 2000 the sequence of the model plant Arabidopsis thaliana. The human genome project involves over a hundred labs, thousands of scientists and eighteen countries contributing significant research (http://www.ornl.gov/hgmis/CENTERS). Starting in 1990 the National Institute of Health (NIH) and the Department of Energy (DOE) planned to spend an average of 200 million US $ per year for fifteen years to complete an amount of sequence data, for which it is estimated that it would require the equivalent of two hundred volumes of the Manhattan phone book to print the result. Anyone attempting to read it would spend the next fifty-two years reading continuously for twelve hours a day (DOE, “Primer on Molecular Genetics” (1992) http://www.bis.med.jhmi.edu). In 1997, the Human Genome Project had around 120Mbp of completed sequence, what represents 2.68% of the human genome, while spending was already 90% of the projected total cost (http://weber.u.washington.edu). In direct competition to these public efforts, two companies (Perkin Elmer and Compaq-Digital), together with Craig Venter founded a new venture, Celera Genomics, to complete the human genome project four years ahead at a total cost of US$250 million. The new company uses capillary systems to automate sequencing (Scherer et al., 1999), which are powerful enough to
sequence 100 million base pairs per day. To process the results, they are building one of the largest computer facilities in the world. Their strategy of shotgun sequencing (Venter et al., 1998) relies on cutting the whole genome into random pieces and then cloning them. Each 10kbp segment will be fragmented and sequenced. By looking for 50bp overlaps, the whole genome will be reassembled. It is estimated that to close the gaps it is necessary to sequence around 35 billion base pairs to assemble the complete genome comprising of 3 billion base pairs.

1.1.1 Comparative gene expression analysis – functional genomics

The increasing amount of sequencing data created by the afore mentioned sequencing projects provide ever more information about the composition of the genomes of many organism. In many cases the function of a novel gene can be assigned on the basis of extensive sequence similarity with known genes from other organisms. However, this approach has its limitation, and still leaves a large fraction of genes with uncertain or unknown function (human: up to 60%). The collection of raw sequencing data has to be supplemented by functional data to offer the possibility of characterising genes in a more functionally oriented way (see 1.3.1.2) The first level of analysis for genome-wide assays is the study of gene expression, for which there are currently two main promising approaches: (1) gene-expression at the mRNA-level (Transcriptomics) or (2) the examination of the proteome (Proteomics).

The tools to characterise mRNA expression levels can be subdivided in three major groups (Kozian and Kirschbaum, 1999): hybridisation-based techniques, PCR-based techniques and sequence-based techniques (see Table 1). The examination of the proteome, in which the expression profile of proteins is analysed, is subject of the main part of this thesis.

1.1.1.1 Hybridisation based techniques – the ‘classical’ approach

A breakthrough in the analysis of gene expression was the development of the northern blot technique in 1977. The expression pattern of mRNA transcripts is studied by hybridising labelled cDNA or RNA to RNA blots (Alwine et al., 1977). The expression of specific RNAs can be detected by Sl-Nuclease/RNase-protection assay. For the RNase-protection assay, the sequence of the analysed mRNA has to be known, since a labelled cDNA strand has to be synthesised to form the RNase resistant RNA-DNA hybrids (Berk and Sharp, 1977).
Table 1: Current methods for the study of gene expression at the transcript level.

To identify specific differences in the expression of cloned cDNAs, differential plaque-filter hybridisation can be used (Maniatis et al., 1978). Although both of these techniques are well established, the major limitation is the applicability only to known genes. Subtractive cloning (Hedrick et al., 1984), has extended the hybridisation method to unknown genes:

SABRE (selective amplification of differentially expressed mRNAs via biotin- and restriction-mediated enrichment) allows the isolation of previously unknown genes. cDNA derived from a tester population is hybridised against the cDNA of a control population. After a purification step specific for tester-cDNA-containing hybrids, tester-tester homohybrids are specifically amplified using an added linker (Lavery et al., 1997).

The most sophisticated hybridisation technique to study gene-expression is done by DNA microarrays (Schena et al., 1998) which allows massive parallel mining of biological data. A microarray consists of many microscopic spots, each of which contains identical single-stranded
Introduction

Chapter 1

oligonucleotides. The application of microarrays fall into two categories: searching for a particular sequence (gene-mutation studies; see 1.3.1.3) or attempting to monitor active gene expression (gene-expression studies). It was possible to show in a gene-expression experiment in yeast, that a whole genome can be analysed in a single reaction (Wodicka et al., 1997). For the manufacturing of microarrays two main techniques are applied. Using the synthesis technology developed by Fodor ((Fodor et al., 1991); Affymetrix, Santa Clara, CA, USA), the microarray is prepared by in situ synthesis from biochemical building blocks in a step-wise fashion. This technique results in what is called a GeneChip™. In this photolithography approach, chips can be directly manufactured using photomasks. The delivery technologies, as the second approach, is using mechanical microspotting developed by Shalon and Brown (Synteni, Fremont, CA, USA) or ink jet technologies developed by Incyte Pharmaceuticals (Palo Alto, CA, USA).

In the mechanical microspotting approach (Scheda et al., 1995) premade biochemical substances are printed directly onto solid surfaces. Using ink jet technology a direct surface contact is not needed and the piezoelectric delivery allows a very high throughput in chip production.

1.1.1.2 PCR-based techniques

Differential display (Liang and Pardee, 1992) of eukaryotic mRNA allows the comparison of the different gene expression pattern of two or more cell populations or tissues. Total mRNA is reverse transcribed and subsequently amplified with random ten-mer primers. The analysis of the radioactive or fluorescence labelled PCR products in a sequencing gel gives information about the gene-expression status of the analysed cell or tissue. To study the difference between two complex genomes Representational Difference Analysis (RDA) takes the advantage of subtractive hybridisation and PCR (Lisitsyn et al., 1993). In a first step, mRNA derived from two different populations, the tester and the control, is reverse transcribed. Following digestion with a frequent cutting restriction enzyme, linkers are ligated to both ends of the cDNA. A PCR step then generates different DNA pools. The linker of the tester and control cDNA are digested and a new linker is ligated to the ends of the tester cDNA. The tester and the control cDNA are then mixed in a 1:100 ratio with an excess of control cDNA to promote hybridisation between single-stranded cDNAs common to both pools. A subsequent PCR step then amplifies only those homoduplexes generated by the tester cDNA, via the priming sites on both ends of the double stranded cDNA. RDA allows a specific amplification of fragments exclusively present in one cDNA pool, owing to an enrichment of sequences unique to the tester.
1.1.1.3 **Sequence-based techniques**

To identify coding regions of a genome two approaches are possible. One can sequence the entire genome, which has now been done with dozens of microorganisms and deduce the putative coding regions using computer algorithms. For eukaryotic organisms, an alternative approach must be used taking into account the non-contiguous organisation of eukaryotic genes. Instead of using an algorithm Adams *et al.* (Adams *et al.*, 1991) decided to use cells to identify active genes by shortcutting the tedious sequencing process. Total mRNA extracted from a tissue is reversed transcribed and the cDNA pool is cloned into a library. Out of the cDNA library, fragments are selected and a short sequence, 400 to 500 base pairs, of each gene is generated. This random sequencing approach generates **Expressed Sequencing Tags (EST)**. Overlapping EST sequences can be assembled into contigs, allowing the complete mRNA sequence of a gene to be determined. By using EST data, it is possible to discriminate between the expressed and non-coding regions of the genome. Another sequenced based approach for the identification of differentially expressed genes is carried out with **Serial Analysis of Gene Expression (SAGE, see Figure 1)**. This method allows the simultaneous analysis of sequences derived from different cell populations or tissues. SAGE is performed in three steps: (1) expressed transcripts are identified by the generation of sequence tags of 10 to 14bp; (2) the ligation of sequence tags generates concatemers that can be cloned and sequenced; and (3) comparison of the sequence data to determine differences in expression of genes that have been identified by the tags. Sequence tags are generated by reverse transcription of mRNA with biotinylated oligo(dT) primer and then digested with a frequent cutting restriction enzyme (the anchoring enzyme). The biotinylated 3’ end of the cDNA strand is affinity purified with streptavidin-coated beads. The obtained cDNA pool is split into two fractions (A and B), to each of which a primer (A’ and B’) that contains a recognition site for a type-IIS restriction enzyme is ligated. After digestion with this tagging enzyme and elution of the digested and unbound DNA portion, the eluted DNA fragments are ligated and amplified using the primers A’ and B’. Following PCR, the primer sites A and B are removed with anchoring enzyme. The sticky ends thus generated enable the DNA fragments to form concatemers, which can then be cloned into a vector. The structure of the concatemers has a typical pattern: between each anchoring site, the so-called ‘digitag’ usually contains the sequence information of two independent cDNA tags. This procedure is performed for every mRNA population to be analysed. Based on the sequence information, comparative computational analysis for the presence and the frequency of transcripts can be performed. In a recent study it was shown that the frequency of genes...
Introduction

Chapter 1

found in SAGE strongly correlates with their abundance in the corresponding cDNA libraries (Velculescu et al., 1995). The major drawback of SAGE lies in the fact, that its efficiency of SAGE is strongly dependent on the completeness of available databases.

The rapid development in the field of protein analysis has the capacity to complement molecular approaches based on DNA and RNA expression studies. Proteomics has the advantage of being closer to the biological consequences of altered gene expression. As an integrative approach it benefits from the wealth of sequence information accumulated by genome-based approaches, which allow peptide data to be connected directly to nucleotide sequences and gene information. The main strength of Proteomics lies not in a massive cataloguing of protein spots, but in allowing proteins to be placed into functional classes by following changes in their expression and post-translational modification levels. Genes are themselves subject to regulation and should be considered as participants in the information flow inside the complex and dynamic processes of a cell. The relationship between genes and proteins is non-linear. The somatic recombination of immunoglobulin genes (Tonegawa, 1983), gene splicing (Berget et al., 1977), RNA editing (Benne et al., 1986) and protein splicing (Kane et al., 1990) makes the one-gene-one-protein hypothesis of Beadle and Tatum (Beadle and Tatum, 1941) no longer valid. To predict protein dynamics using genomic data is not possible, because there is no reliable correlation between levels of gene transcription and protein abundance (Anderson and Seilhamer, 1997). This has also been shown by studies by Aebersold’s group (Gygi et al., 1999). It was found that the correlation between mRNA and protein levels was insufficient to predict protein expression levels from quantitative mRNA data. Indeed, for some genes, while the mRNA levels were of the same value the protein levels varied by more than 20-fold. Conversely, invariant steady-state levels of certain proteins were observed with respective mRNA transcript levels that varied by as much as 30-fold. The existence of alternative splicing of mRNA, post-translational modifications of proteins, the different way of processing of proteins under different conditions and the final degradation makes it impossible to automatically correlate mRNA and protein abundance in a cell.
Chapter 1

Introduction

Figure 1: Experimental procedure for the serial analysis of gene expression (SAGE) (Velculescu et al., 1995).
"The proteome refers to the total set of proteins expressed in a given cell at a given time."

MARC WILKINS (1994)

1.2 Proteomics

Proteomics (James, 1997) is a field that promises to bridge the gap between genome sequence and cellular behaviour (see Figure 2). It aims to study the dynamics of the protein products of the genome and their interactions, rather than focusing on the simple semi-static DNA blueprint of a cell. Genome sequencing is providing a huge amount of data on predicted gene products, but the majority of these have no known function. To elucidate the gene function, Proteomics analyses the proteins synthesised under a given set of conditions. While genomics provides an overview of the complete set of genetic instructions available to a cell, functional genomics has come to mean the study of expression and behaviour of gene products (Dove, 1999). This field is now subdivided to focus on particular types of gene products. Transcriptomics compares mRNA expression profiles for example using DNA microarrays (see 1.1.1). Displaying the cells’ protein by electrophoretic methods, analysing the proteins by spectrometric or chemical methods and the identification and interpretation of the obtained data by informatic methods is the realm of Proteomics. Proteomics in turn can be subdivided (Blackstock and Weir, 1999) into Functional Proteomics, the study of global changes in protein expression, and Cell-map Proteomics, the systematic study of protein-protein interactions through the isolation of protein complexes. Metabolomics describes the final step in understanding an organism’s entire metabolism (Oliver et al., 1998).

Since O’Farrel (O’Farrel, 1975), Klose (Klose, 1975) and Scheele (Scheele, 1975) demonstrated that it was possible to separate proteins based on their isoelectric points and molecular weights by electrophoresis on polyacrylamide gels, two-dimensional gel electrophoresis has remained unchallenged as the most efficient way of separating, visualising and quantifying complex protein mixtures. Gels are now highly reproducible using immobilised pI gradients in the 1st dimension, and Coomassie® Blue stain or silver staining allow protein quantitation. Using fluorescence dyes the visualisation of thousands of proteins quantitatively over a wide dynamic range is now possible. The analysis of protein spots is either carried out by traditional techniques as western blotting or classical Edman sequencing or by using a combination be-
tween chemical degradation and high-sensitivity mass-spectrometry methods. The identification of a protein by peptide mass fingerprinting (PMF), peptide fragment fingerprinting (PFF) and supporting chemical methods is well developed. The bottleneck of Proteomics is now more a problem of interpreting the data obtained and locating the protein or cognate EST in a database. To apply the technologies mentioned above, two related but distinct ways can be discerned. First, Functional Proteomics, which refers to the creation of quantitative maps of protein expression from cell or tissue extracts, akin to the EST maps commercially available. This approach relies on 2D-gel maps and image analysis. It opens up the possibility of studying cellular pathways and their perturbation by disease, drug action or other biological stimuli at the global proteome level, thereby offering the potential to find disease markers and elucidate biological pathways. Second, cell-map Proteomics, which deals with the determination of proteins, of their subcellular location and of protein-protein interactions. This approach aims on the identification of cellular machines (Alberts and Miake-Lye, 1992) built by complexes of proteins which perform cellular functions cooperatively. Systematic identification of protein complexes or protein-protein interactions would permit these machines to be defined and allow physical maps to be created for a variety of cell types and states. Such information is of great value for the assignment of protein function, the principal problem in the post-genome era (see 1.3.1.2).

1.2.1 Technology for Proteomics

Figure 2: The taxonomy of genomic biology and flow-chart of thesis.
1.2.1.1 Separating and visualising proteins

In combination with a complete genome sequence, the proteome approach could, in theory, be used to identify every protein on a two-dimensional gel, allowing meaningful comparisons between different cells or physiological states. In practice, however, several limitations remain in the sensitivity, reproducibility and analytical power of Proteomics. Beside systematic biases that exclude very small, very large, very basic and membrane bound proteins, two-dimensional electrophoresis also fail to resolve the proteome completely. Getting reproducible results is also difficult. To improve resolution and reproducibility, large format gels measuring up to 40 x 40cm can be run using immobilised isoelectric gradients. Protein labelling by fluorescent dyes overcomes some of the drawbacks of Coomassie® and silver staining methods and recent progress in automation of gel handling and image analysis speeds up the proteome approach. However, the problem of visualising low-copy-number proteins (1-1000 copies per cell) remains, and housekeeping proteins (>10000 copies per cell) can easily obscure minor components. A prefraccionation by centrifugation or free-flow electrophoresis is one way to overcome these limitations. Special procedures for very basic proteins, membrane proteins and other poorly soluble proteins are also currently under development. One of the advantages of 2D-based Proteomics is the separation of post-translationally or proteolytical modified proteins from the parent protein. It has been suggested (Rabilloud et al., 1997), that up to a quarter of the spots on a gel may be modified proteins and it is not uncommon to isolate what is essentially the same protein in several places on a gel (Paola Dainese-Hatt, personal communication). Although recent improvements in automated spot detection and excision are being made, there are still big limitation in the speed to process gels. The commercial software packages are all designed for analysing a few gels per day and can require 1-8h additional manual editing per gel. This is an impossible limitation when 200-400 gels per week should be analysed and if the software is to be used to drive a spot-excision robot. There is currently a need for better image analysis software packages and full integrative data-storage programs, that can be linked with subsequent robotics to allow gels to be imaged and spots excised and characterised in one unattended operation.

For a detailed introduction and application of Separating and Visualising Proteins, see chapter 2.
1.2.1.2 Detecting proteins

Proteins displayed by electrophoretic methods can be analysed by a wide range of methods. Comigration and HPLC mapping was one of first analysis methods on 2D gels. This method was used to analyse proteins isolated from beef, pork, chicken and Soya bean (Medina and Phillips, 1982). The protein extracts were heated, enzymatically hydrolysed and subsequently analysed by TLC and HPLC. The statistical analysis of the chromatographic digest pattern revealed for example that an all-beef frankfurter can be discriminated from a standard frankfurter containing 35% pork protein. Proteins can also be detected using immunoblotting. Starting from a known protein sequence, every immunogenic peptide can be synthesised, purified, ligated to a hapten and finally an antibody raised against the epitope of choice. As an alternative to peptide synthesis, a phage display system could be used to raise antibodies against the protein of choice.

For a detailed introduction and application of Detecting Proteins, see chapter 3.

<table>
<thead>
<tr>
<th>PROTEIN ATTRIBUTE</th>
<th>SOURCE OF ANALYTICAL DATA</th>
<th>ANALYSIS TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species of origin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isoelectric point</td>
<td>Gel image analysis</td>
<td>Parallel</td>
</tr>
<tr>
<td>Apparent mass</td>
<td>Gel image analysis</td>
<td>Parallel</td>
</tr>
<tr>
<td>Mass</td>
<td>MALDI-TOF MS</td>
<td>Serial</td>
</tr>
<tr>
<td>Protein N- and C-terminal sequence tag</td>
<td>Chemical sequencing of proteins immobilised on membranes</td>
<td>Parallel</td>
</tr>
<tr>
<td>Edman sequencing</td>
<td>Chemical sequencing of proteins immobilised on membranes</td>
<td>Serial</td>
</tr>
<tr>
<td>Peptide mass fingerprinting (PMF)</td>
<td>MALDI-TOF MS or ESI MS of peptide digests</td>
<td>Serial</td>
</tr>
<tr>
<td>Peptide fragment fingerprinting (PFF)</td>
<td>Fragmentation of peptides from PMF and MS/MS analysis</td>
<td>Serial</td>
</tr>
<tr>
<td>Peptide ladder sequencing (PLS)</td>
<td>Mass difference of sequential degradation products</td>
<td>Parallel</td>
</tr>
<tr>
<td>de novo MS sequencing</td>
<td>Fragmentation of peptides from PMF and ion trap MS</td>
<td>Serial</td>
</tr>
<tr>
<td>Amino acid composition</td>
<td>Chromatographic analysis of hydrolysed proteins</td>
<td>Serial</td>
</tr>
</tbody>
</table>

*Table 2: Primary and secondary protein attributes; modified after (Wilkins and Gooley, 1997).*
Today most strategies of protein identification involve the definition of one or more attributes of proteins, which are then matched against databases in various manners. A primary attribute of a protein is a property of or generated directly from the intact protein. A secondary if it represents or is generated from fragments of the whole molecule (see Table 2; (Wilkins and Gooley, 1997).

Mass spectrometry is now the method of choice for fast protein detection and offers the possibility to determine post-translational modifications. Because isoelectric point and molecular weight are insufficient to identify a protein, the spot of interest is digested by a protease and the resulting peptides subsequently mass analysed. A short introduction into the principles of mass spectrometry is given and its application as a useful laboratory tool is shown. The use of mass spectrometry for peptide sequencing or in combination with computer algorithm is subject of the following chapters.

For a detailed introduction and application of DETECTING PROTEINS, see chapter 3.

1.2.1.3 Identifying proteins

The historically most popular method to identify a protein is N-terminal sequencing using the chemistry defined (Edman, 1949) and automated (Edman and Begg, 1967) by Pehr Edman. Automated Edman sequencing is a useful method to generate N-Termini for primer construction, easy to use and the sequencing results not complicated to interpret. Full sequencing of a protein is usually not done due to the long cycle times, costs and the relative high amount of material needed.

Sequencing by mass spectrometry is as well possible by using Matrix-Assisted-Laser-Desorption and Ionisation (MALDI) Time-Of-Flight (TOF) Mass-Spectrometry (MS) Post Source Decay (PSD) analysis (Kaufmann et al., 1993) or ESI (Electrospray Ionisation) MS/MS (Hunt et al., 1981). Both spectrometric methods, MALDI-TOF and ESI, are suitable for automation and can detect peptides at the femtomolar level. To realistically identify a protein spot an amount in the picomolar range is needed. This corresponds to a protein present in 1000 copies per cell isolated from 10^3-10^9 cells. With the automatisation of spot excision, proteolysis and peptide extraction, MALDI-TOF has the potential to identify several thousand peptides a week, whereas ESI although with lower throughput, offers additional information on partial sequence and on post-translational modifications. In a normal experiment, the methods are used hierarchically. Although there is a way of sequencing a protein using a reflectron
MALDI mass spectrometer by PSD analysis, which can yield partial sequence information, the method of choice for peptide fragmentation and sequencing is currently Collision-Induced-Fragmentation (CID). These kinds of experiments are carried out with instruments equipped with a collision cell, such as an ion trap or a triple quadrupole instrument. To ease the interpretation of CID data several chemical methods are used. The digestion in a mixture of $^{16}\text{O}$ and $^{18}\text{O}$-labelled water, results in a specific label of the C-Termini (y-ion series). The modification on the N-Termini by H$_4$/D$_4$-Nicotinic Acid-NHS esters results in a specific label of the b-ion series.

For a detailed introduction and application of IDENTIFYING PROTEINS, see chapter 4.

1.2.1.4 Processing peptide data

MALDI TOF is usually used for Peptide-Mass-Fingerprinting (PMF; (Henzel et al., 1993) (James et al., 1993) (Mann and Wilm, 1994) (Pappin et al., 1993) (Yates et al., 1993). The masses of peptides derived from on-blot or in-gel proteolytic digested proteins are measured and searched against a computer-generated digest of protein databases using the same enzyme. With high mass accuracy measurement (below 10 PPM) and available data of completely sequenced organisms (e.g. Caenorhabditis elegans or Saccharomyces cerevisiae) this technique alone is sufficient to identify proteins (Shevchenko et al., 1996). ESI coupled to a quadrupole, ion trap or a time-of-flight analyser is the method of choice, if full-sequence information is not available to generate additional partial sequence information to complement the peptide mass information. The obtained peptide sequence tags (Mann and Wilm, 1994) can be used to search protein and EST databases. A different approach uses a correlation algorithm developed by Yates’ group to match uninterpreted tandem mass spectra with computer-generated mass spectra from protein and EST databases (Peptide Fragment Fingerprinting, PFF; (Yates et al., 1995). In Peptide Ladder Sequencing (PLS), the sequence is read by the mass differences between sequential degradation products using MALDI-TOF MS (Chait et al., 1993, Korosten-skyy, 1998 #93).

With the rapid growth of nucleic acid and protein databases, there is increasing concern about the impact of this information explosion (Bork and Koonin, 1998). The eventual success of genome projects will depend on the ability to handle information in a manner that enhances the capability for function prediction. Improvements in the sensitivity of mass spectrometry and innovation in the instrumentation such as z-sprays and quadrupole time-of-flights (Q-TOF) will facilitate data interpretation. The more immediate gains will come from the improvement of
search algorithm and integrated software (see Table 3), using bioinformatic tools to handle the huge amount of data generated by high-throughput genomics (Blackstock and Weir, 1999). Additional improvement can be achieved by optimising the signal-to-noise ratio by the reducing chemical noise generated from all the methods applied upstream to data acquisition. Better gel material, clean rooms and automation in sample handling (reducing the popular keratin pollution) will improve the quality of data and its interpretation. Even approaches to skip the separation of proteins by two-dimensional gel electrophoresis by using multidimensional chromatography and tandem mass-spectrometry are under development.

For a detailed introduction and application of PROCESSING PEPTIDE DATA, see chapter 5.

<table>
<thead>
<tr>
<th>TYPE OF PROGRAM</th>
<th>NAME</th>
<th>URL ADDRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intact protein molecular mass and</td>
<td>PeptideSearch</td>
<td><a href="http://www.mann.embl-heidelberg.de/Services/PeptideSearch.html">http://www.mann.embl-heidelberg.de/Services/PeptideSearch.html</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MassSearch</td>
<td><a href="http://www.chrg.inf.ethz.ch/subsection3_1_3.html">http://www.chrg.inf.ethz.ch/subsection3_1_3.html</a></td>
</tr>
<tr>
<td></td>
<td>MS-Fit</td>
<td><a href="http://falcon.ludwig.ucl.ac.uk/MS-Fit.html">http://falcon.ludwig.ucl.ac.uk/MS-Fit.html</a></td>
</tr>
<tr>
<td></td>
<td>PeptideSearch</td>
<td><a href="http://www.mann.embl-heidelberg.de/Services/PeptideSearchForm.html">http://www.mann.embl-heidelberg.de/Services/PeptideSearchForm.html</a></td>
</tr>
<tr>
<td></td>
<td>ProFound</td>
<td><a href="http://chait-sgi.rockefeller.edu/cgi-bin/prot-id.html">http://chait-sgi.rockefeller.edu/cgi-bin/prot-id.html</a></td>
</tr>
<tr>
<td>Peptide Fragment Fingerprinting</td>
<td>Mowse</td>
<td><a href="http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse">http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse</a></td>
</tr>
<tr>
<td></td>
<td>MS-Tag</td>
<td><a href="http://falcon.ludwig.ucl.ac.uk/mstag.html">http://falcon.ludwig.ucl.ac.uk/mstag.html</a></td>
</tr>
<tr>
<td></td>
<td>PepFrag</td>
<td><a href="http://chait-sgi.rockefeller.edu/cgi-bin/prot-id-frag">http://chait-sgi.rockefeller.edu/cgi-bin/prot-id-frag</a></td>
</tr>
<tr>
<td></td>
<td>PeptideSearch</td>
<td><a href="http://www.mann.embl-heidelberg.de/Services/PeptideSearch/FR_PeptidePatternForm.html">http://www.mann.embl-heidelberg.de/Services/PeptideSearch/FR_PeptidePatternForm.html</a></td>
</tr>
<tr>
<td></td>
<td>Sequest</td>
<td><a href="http://thompson.mbt.washington.edu/sequest.html">http://thompson.mbt.washington.edu/sequest.html</a></td>
</tr>
<tr>
<td>Peptide Ladder Sequencing</td>
<td>MassDynSearch</td>
<td><a href="http://cbig.inf.ethz.ch">http://cbig.inf.ethz.ch</a></td>
</tr>
<tr>
<td></td>
<td>MS-Edman</td>
<td><a href="http://falcon.ludwig.ucl.ac.uk/mstag.html">http://falcon.ludwig.ucl.ac.uk/mstag.html</a></td>
</tr>
<tr>
<td></td>
<td>Mowse</td>
<td><a href="http://serv1.dl.ac.uk/SEQNET/mowse.html">http://serv1.dl.ac.uk/SEQNET/mowse.html</a></td>
</tr>
</tbody>
</table>

Table 3: Protein identification and characterisation programs using MS data. Modified table from (Wilkins and Gooley, 1997).
1.2.2 Application of the technology

As pointed out in the previous sections, proteomics integrates and optimises analytical techniques from the whole life science field with respect to high-throughput, speed and sensitivity. Proteomics aims to reveal all proteins present under certain conditions. To stretch the meaning on the functional aim, one speaks today of functional proteomics. The strategy of functional proteomics is to uncover the action and interaction of all the proteins in a given organism. But there must be as well an hierarchical approach, because the variation and differences in protein function provides a way to group proteins in functional related groups creating a map of all the interactions inside the cellular machinery.

1.2.2.1 Functional Proteomics

The quantitative analysis of cells or tissues in different physiological states using 2D gels and image analysis has the advantage of simultaneously monitoring the complex behaviour of physiological changes at the protein level. A further asset is the possibility to detect post-translational modifications as glycosylation.

The Hochstrasser Lab in Geneva is using proteomics as a clinical diagnostic tool. The search tools available at the SWISS-PROT database (http://www.expasy.ch) are especially useful to study disease processes. Celis and co-workers are using a large human keratinocyte database in the study of bladder squamous cell carcinomas. The construction of a 2D database with the help of bioinformatics has been instrumental to identify markers that may differentiate various histopathological biomarkers (Rasmussen et al., 1996). In a systematical study, they identified a protein called psoriasin in urine samples of patients with squamous cell carcinoma of the bladder. It is now possible to use psoriasin as an early marker of the disease. Mose Larsen and his group compared normal and β–cells from the islet of Langerhans to search for changes in insulin-dependent diabetes mellitus (Aanstoot et al., 1996). In a correlation study between mRNA levels and detectable protein levels a poor correlation was found (Anderson and Seilhamer, 1997). The protein-expression monitored by 2D gel analysis and mRNA-levels obtained by transcript imaging (see Table 1) did not show a 1:1 ratio, implying that for accurate quantitative work it might be preferable to work at the protein level.

Proteomics is not only more direct than genomics, it also offers the capability to bring a rapid focus to more than a handful molecular targets. Oxford GlycoScience (Abingdon, UK) completed the first phase of a proteomic study of disease-specific proteins in the synovial fluid of
patients with rheumatoid arthritis. From an initial sample of over 500 proteins, 47 were identified as disease specific. 12 of the disease specific proteins identified showed no known correlation to proteins in humans and serve now as a starting point for further studies (Parekh, 1999).

### 1.2.2.2 Cell-map Proteomics

While functional proteomics has a global view of the complex network of proteins, cell-map proteomics looks at a defined subcellular complex, subdividing the proteome in little interdependent compartments, so revealing functionally distinct cellular machines (Alberts and Mikael-Lye, 1992).

A great deal of information can be acquired about the function of a protein if one looks at its partner, its location and its behaviour upon changes in the environment of the cell. The measurement of protein-protein interactions by purification of protein complexes and subsequent mass-spectrometric analysis has elucidated the components of the spliceosome complex (Neubauer et al., 1997). A systematic high-throughput approach proposed by Kuster and Mann (Kuster and Mann, 1998) for the characterisation of protein complexes implies gene tagging with peptide sequences that enable affinity purification, subcellular fractionation, affinity purification and mass spectrometry. A similar approach is already undertaken by using the two-hybrid system for all the 6200 proteins of the yeast proteome (Fromont-Racine et al., 1997). This technique identifies pairwise interaction rather than the whole complexes and relies on proteins entering and folding in the nucleus.
1.3 Economics

Drug companies are fond of talking about their ‘pipelines’ – the range of compounds they have in various stages of development in their laboratories, ready to take over as money-making machines when the patents on their existing product expires (Carr, 1998). These pipelines, however, are pretty leaky affairs. For every approved drug that comes out of a pipeline, about 10,000 molecules have gone in and got lost somewhere on the way. The molecules put into the funnel are the company’s raw material (see Figure 3). They are stored in libraries, which are generated by combinatorial chemistry or are the natural products of living organisms. The traditional process of drug discovery starts by screening a library in tissue cultures or by enzyme assays to see if any of the compounds it contains have properties that could be used to treat all kind of diseases. Substances emerging from this initial screening are rarely powerful enough to be effective, so the next step is to optimise the biological activity by rearranging the compound using chemical, biochemical and molecular biological techniques. The lead compound, which results from these initial steps, is further subjected to activity tests in animals and various tissue cultures. These safety studies show the bioavailability, the toxicity and possible side effects to be expected. Only then the compound is allowed to go into clinical trials in a small number of people (phase 1). After passing through this stage, the compound has to prove its effectiveness and safety in a larger number of patients. After passing phase 3 of clinical trials the surviving compound/company has to persuade the regulatory authorities to agree to releasing the compound out on the market. Compounds leak out of the pipeline at every stage. Even after passing the initial screening and the clinical test trials, only one molecule in ten makes all the way through. The average cost of bringing a compound to the market is now 300 $m, that is distributed on 200 $m in pre-clinical and 100 $m in clinical trials. Viewed from the laboratory, a pipeline is like a funnel, getting narrower towards the end as fewer compounds are left in. In the view of a financial department, it looks more like two funnels joined together. Expenditures contracts towards the middle as compounds are getting lost, then increasing again as the few candidates left are put through successively larger and thus more expensive clinical trials. The
period of patent protection starts when the compound is registered with the patent office, not when it first appears on the pharmacist's shelves, which can be much later. So drug companies have only a few years to recoup their R&D costs and to earn profits. For an average drug, every delay after patent has been applied costs $\text{18m} \text{ per day in protected sales}$. Therefore the pharma industry has three wishes: to produce more effective drugs for a wider range of diseases, to make research and development less expensive (see 1.3.1) and to speed it up so that they can benefit from patent protection for longer times (see 1.3.2).

![Diagram of drug development in years](image)

**Figure 3:** Drug development in the pharmaceutical industry (Carr, 1998); the size of the funnel reflects the number of drug candidates.

### 1.3.1 More effective drugs

To produce more effective drugs for a wider range of disease, biotechnology companies have worked out a way of finding new drugs without screening libraries in a traditional way. Companies invest a huge amount of money to get access to databases of genomic sequences provided by Incyte or Human Genome Sciences. What all of these companies are buying into is a developing model of how organisms including humans work (Friedrich, 1996). Medical drugs can be broadly divided into two classes. The first acts directly on the metabolism, the second on invading pathogens or parasites. Both kinds of drugs act by specific chemical reactions; the more specific the better. Of the 3000 metabolic medicines sold today, 85% are targeted on 417 molecules. For pathogens and parasites, there are only 66 known targets (Waters, 1997). An effective drug has only one single molecular target. The broader the range of the drug's interaction with other molecules, the more side effect the drug causes. Most FDA approved drugs
show a common pattern: they work for most users, but they are ineffective or even harmful to some individuals. Some drugs produce side effects in up to 40% of patients in treatment (Fisher, 1998). To develop new drugs against current untreatable diseases the identification of new target drugs is necessary. The databases of genomic sequences offer new ways to identify these new targets. Because 45% and 5% of the already known drugs act on the family of the 7-transmembrane receptors and ion channels respectively, databases are screened for these kind of motifs and new potential genes are cloned \textit{in silico} (Carr, 1998). Another way of identification would be the comparison of the tissue or state specific proteins of a healthy cell with a diseased cell by subtraction of the normal expressed proteins to identify the tissue or state specific proteins. For this purpose proteomic databases giving the information of the dynamic expression state of a cell are needed. A way of finding new parasite drugs is done by sequencing the parasite of choice, which can be done in theory in a week by using the latest capillary sequencer and shotgun techniques (Venter et al., 1998). The comparison of the gene sequences with genes with known functions is giving hints for possible new drug targets.

1.3.1.1 Pharmacogenomics

Pharmacogenomics (Kleyn and Vesell, 1999), which is the elucidation of drug action based on genotype and expression profile of RNA or proteins, aims to discover the molecular efficiency and toxicity markers that will aid in the development of novel, more effective and safer drugs. Pharmacogenomics applies the large scale systematic approaches of functional genomics to speed the discovery of drug response. The hypothesis underlying pharmacogenomics is that often complex diseases as cancer, schizophrenia or hypertension are the result of several environmental and genetic factors leading to a similar clinical outcome.

1.3.1.2 Pharmacogenomic technologies

Pharmacogenomic studies use three kinds of technologies: first, genome-wide DNA genotyping to correlate the presence of an allele with a trait (disease or drug response) in unrelated individuals of a population. To conduce these so-called association studies single nucleotide polymorphism (SNP) are used. SNPs are simple base-pair substitutions that occur within and outside genes. The success of this approach depends on the population-specific relative risk associated with each drug response allele. Alleles with a strong relative risk such as APOE 4 for Alzheimer’s disease (AD) can be detected with genome-wide DNA genotyping (see (Zubenko et al., 1998) and 1.3.1.3). The phenotypic consequence of polymorphism affecting
drug response are presumably associated with changes at the gene expression level. Transcriptomics (see Table 1) and Proteomics (see 1.2) are the methods of choice to be used for pharmacogenomic studies.

1.3.1.3 Pharmacogenomic developments

Many drugs are causing enormous side effects in a high percentage of patients. The treatment of such diseases is often a hit and miss affair. For example Clozaril™ from Novartis induces agranular cytosis in 2% of schizophrenia patients treated with. In Parkinson’s disease, a range of different drugs are needed to combat the disease, because only one or two of them will work in a given patient. Hypertension can be treated for example in four different ways: ACE inhibitors, beta-blockers, calcium channel blockers and diuretics. Matching the treatment to a particular patient is so far a trial and error matter. Knowing the alleles of a patient makes it possible to choose the right medication. Myriad Genetics (Salt Lake City, UT, USA) is developing a test for mutations of the angiotensinogen (AGT) gene, which encodes a protein that regulates salt retention. Genetic profiling by correlating AGT mutations with drug response, will allow physicians to choose hypertension treatments more rationally (Marshall, 1997). Alzheimer disease (AD), whose causes are still mysterious, is known to be somehow related to APOE-genes. Pharmacogenomics is trying to identify specific genotypes and customise drugs to subgroups. This will lead to several potential products: personalised genetic ID cards or personalised drugs (Marshall, 1997). Pharmacogenomic studies on the Alzheimer drugs provided by Servier (Lille, France) and Park-Davis (Ann Arbour, MI, USA) showed that Servier’s drug (designated S12024, which is thought to increase brain noradrenergic/vasopressinergic activity) works best in patients homozygotic for the APOE-ε4 allele (ε4/ε4) (Poirier et al., 1995). ParkeDavis’s drug Cognex™ (tacrine hydrochloride; cholinesterase inhibitor, which is thought to act by raising levels of acetylcholine, profoundly depleted in brains of AD patients) conversely showed better results in APOE-ε2 and APOE-ε3 patients (Richard et al., 1997). Pharmacogenomics using proteome and genome tools will allow companies to aim existing drugs more precisely and even let them rehabilitate drugs, which leaked out of the pipeline due to side effects in clinical trials. Another interesting area to extend a company’s portfolio will be the area of diagnosis. This term, describing the detection of disease, is estimated to generate a US$7 billion business. Rapid genotyping of bacteria or viruses for drug resistance will allow physicians to prescribe the correct treatment. GlaxoWelcome (London, UK) is working with Affymetrix (Santa Clara, CA, USA) to develop a DNA chip (see 1.1.1.1) to screen for HIV strains with variant rev genes that are resistant to reverse transcriptase inhibitors. Francis Collins de-
Chapter 1 Introduction

scribes the importance of these tests: “With the ability to test for disease genes, we can design medical programs for individuals that include lifestyle, diet and medical surveillance to alleviate or prevent disease” (Collins, 1995). The pharmaceutical companies promise three things (Carr, 1998): first, an increase in the range of diseases that are treatable with drugs; second an increase in the precision and effectiveness of those drugs; and third, an increase in the ability to prevent diseases, rather than just treating it.

1.3.2 Increasing research effectiveness

Pharmaceutical companies are at present looking to high throughput screening (HTS) and combinatorial chemistry to feed their pipelines. It is estimated that the top 10 global pharma companies will require five new chemical entities (NCE) per year, each with a salary potential of US$350 million per year (the so-called “blockbuster”) in order to maintain their current growth rate of 10%. The launch of five NCEs a year is estimated to need to select 15-30 compounds for clinical development and a must to complete 35 to 75 discovery projects a year (Streng and Lloyd, 1998). A direct result of this requirement will be the formation of external alliances and a decrease in the development time needed to bring a compound from the lab bench to the pharmacy counter. One way of doing this is using external knowledge of highly specialised companies e.g. in combinatorial chemistry and finding and feeding the pipeline with outside developed promising targets (AXYS Pharmaceuticals, San Francisco). Covance, (Princeton, NY, USA) is another of such so-called contract-research organisations (CRO). This company is running clinical trials by networking computing and simultaneous integration of clinical data obtained. They provide pharmaceutical companies a way to conduct faster and cheaper trials and extract the maximum number of information from the minimum number of patients needed. A growing number of drugs are going in the post approval phase-IV trials (the so-called ‘pushing the edge of the envelope phase’), in which the range of applicability of a drug is extended.
1.4 References


2 Separating and Visualising Proteins

The first and most fundamental part of a Proteome project comprises the separation and visualisation of proteins present in a cell under certain conditions. The solubilisation of the whole protein content of the cell must be achieved, which is crucial for all the subsequent steps described in the following chapters. If a protein is sparsely soluble in the solubilisation buffers applied, it will tend to denature and precipitate. There will be no chance to detect it, even with the most advanced mass spectrometric techniques available. In the next step, the cell proteins have to be separated from each other in a reproducible way. This is also far from being trivial, because many hydrophobic proteins even if solubilised stick together at their isoelectric point and tend to form vertical smears in the second dimension. To separate a complex protein mixture ("the Holy Grail of two-dimensional electrophoresis", Denis F. Hochstrasser) at least two orthogonal separation methods have to be applied. The most common are isoelectric point and mass, but every other combination is possible, when it is linked to the intrinsic nature of a protein like hydrophobicity or buoyancy factor. The third crucial point is the visualisation of all the proteins solubilised and separated. This is of great importance, especially if one aims to quantitate the proteins displayed or to compare different effects of the expression of different proteins. This implies that the dye staining should be reproducible, linear with quantity and independent of the nature of the protein.

If detergents are chosen, which are incompatible with subsequent analysis steps, most of the analyses just do not work at all. It would be nice to present a universal separating and visualisation procedure, but the method presented is only optimised for the processing of the soluble protein content of Bradyrhizobium japonicum. Even changing to closely related strains as Sinorhizobium melliloti or Rhizobium leguminosarum, the method has to be changed and optimised again.

This chapter will give an introduction to two-dimensional gel electrophoresis and discuss different strategies of visualisation. An overall view of the complex heat-shock regulatory network of B. japonicum will be used to illustrate these strategies. In the following chapters, this regulatory model will serve as an example for the application of the methods described. As an example for subtractive gel-analysis, the two dimensional electrophoresis of a HrcA mutant is shown. The analysis confirmed the role of HrcA as a negative control element in the regulatory heat shock network of B. japonicum. Other examples for the application of two-dimensional electrophoresis can be found in:


Santella, L., Kyozuka, K., Hoving, S., Münchbach, M., Quadroni, M., James, P. and Carafoli, E. Breakdown of cytoskeletal proteins and calpain action during meiosis resumption by starfish oocytes. Submitted to Development.

2.1 Separating Proteins

The separation of proteins implies the solubilisation and resolution of a complex protein mixture. Today, the most popular way for the analysis of the complete complement of proteins is done by two-dimensional (2D) polyacrylamide gel (PAGE) analysis ("A two-dimensional gel is like stars in the sky", Keith L. Williams).

The 2D-PAGE technique which is mostly used today originated from the work of Patrick O'Farrell (O'Farrell, 1975), G. Scheele (Scheele, 1975) and Joachim Klose (Klose, 1975). The technique became popular because of its tremendous resolving power and sensitivity as well as the ability of being compatibility with other methods such as electrophoretic blotting, Edman sequencing or mass spectrometry. It can be scaled up and processed in a highly parallel manner and even automated. 2D PAGE is based on the separation of proteins according to their isoelectric points, pl, in the first dimension and by mass in the second-dimension SDS gel electrophoresis.

2.1.1 Sample preparation

Pretreatment of samples for isoelectric focusing (IEF) involves solubilisation, denaturation and reduction of the sample. Non-protein components such as lipids or nucleic acids (Rabilloud, 1996) must be eliminated from the samples. All the reagents used must be compatible to the subsequent isoelectric focusing step. The challenge for 2D-PAGE is the solubilisation and separation of hard to solubilise proteins in a way compatible to the isoelectric focusing step. The critical point is that proteins have to be fully denatured – to display all their charged groups – but at the same time remain in solution. This is a highly unnatural state. Extreme care has to be taken, not to disturb the natural net charge of a protein by detergents as sodium dodecyl sulphate. Rabilloud (Rabilloud et al., 1997) reported the selective adsorption of hydrophobic membrane proteins to the IPG during the IEF, resulting in these proteins being under-represented on the second dimensional gel. It was shown that the incorporation of thiourea or propanesulphonates (e.g. CHAPS) improved the solubilisation. Removal of nucleic acids is another critical point in the sample preparation step. DNA complexes dissociate under denaturing conditions, causing a marked increase in viscosity of the solution, inhibiting protein entry into the first dimensional gel. DNA binds as well to proteins in the sample and causes artificial migration behaviour (Rabilloud, 1996) and can as well cause smearing effects on the
second dimension. Therefore DNA and RNA are removed by RNase/DNase treatment and by the addition of ampholines to complex them.

2.1.2 The first dimension

In the traditional way of separating proteins by IEF, the pH gradient is maintained dynamically with carrier ampholines as introduced by O'Farrell. Using carrier ampholines it was difficult to achieve good reproducibility even in the same laboratory. A second and even more frustrating thing was the inability to separate preparative loads of proteins, to allow acquiring a reasonable amount of material in a reasonable number of experiments. The introduction of immobilised pH gradients (IPG) eliminated the problems of gradient instability and small loading capacity (Bjellqvist et al., 1982). To form an IPG, acrylamide is derivatised with groups displaying different pI and subsequently co-polymerised with the acrylamide gel matrix, which results in completely stable gradients. The availability of commercially precasted gels of different pI ranges gives reproducible first dimensional gels. Nowadays even gram quantities of protein for micropreparative purposes can be loaded on the first dimension gel. However, the technique is still far from being perfect. The separation of basic (pI > 10) and very acidic (pI < 3) proteins on IPGs is so far not possible on a regular base due to the instability of the gel matrix at high and very low pH. This problem is of great importance, because the majority of DNA-binding proteins as transcription factors or regulatory elements have a high intrinsic basicity. Another major problem is the displaying of low abundance proteins. Low abundance proteins often have important regulatory roles, but many will remain undetected on broad pH range micropreparative 2D-PAGE gels. Even when high quantities are loaded on a gel, the low abundant protein of interest might be covered or contaminated by neighbouring more abundant protein. To address this problem prefractionation of the sample has to be done and additional steps before loading on the gel are necessary.

2.1.3 The second dimension

In the second dimension, several different gel systems can be applied. The most common is a large version of up to 25cm x 25cm of a standard SDS-PAGE (Laemmli, 1970). Nevertheless, the theme can be varied, using Tris-Tricine gels (Schagger and von Jagow, 1987), the BAC-gel system (Hartinger et al., 1996), acidic gels (Kaltschmidt and Wittmann, 1970) or gradient gels in all variations.
2.1.4 Carrying out two dimensional gel electrophoresis

2.1.4.1 Materials

Bulk chemicals:

- NaCl (Baker 0278); Trizma base (Tris, Sigma T1503); MgCl₂ (Fluka 63064); ethanol (Fluka 02860); urea (Fluka 51456, MicroSelect); sodium dodecyl sulphate (Fluka 71729); 1,4-dithio-DL-threitol (Fluka 43817); iodoacetamide (Fluka 57670); 3-[(3-cholamidopropyl)dimethyl-ammonio]-1-propanesulfonate (CHAPS, Sigma C3023); glycerol (Fluka 49782); silicone oil DC 200 (Fluka 85412); glycine (Fluka 50050); methanol (Lab-Scan, HPLC pure, C2517); acetic acid (Scharlau, Ac 0344)

- Enzymes:
  - DNase I (Boehringer Mannheim 104 159); RNase A (Boehringer Mannheim 109 169)

Electrophoresis:

- Acrylamide (BDH Electran® 44299 4J); NN'-methylenebisacrylamide (BDH Electran® 44300 3N); ammonium peroxodisulfate (Fluka MicroSelect 09913); sodium dodecyl sulphate (BDH 442442F); N,N,N',N'-tetramethylenediamine (Fluka 87689)

Ampholines/Immobilines:

- Resolyte® pH 4 -8 (BDH 44340 3C); Resolyte® pH 3.5 - 10 (BDH 44338 3P); Pharmalyte® pl 8-10.5 (Pharmacia 17-0455-01); Immobiline™ DryStrips (pH 3-10L, 18cm; Amersham Pharmacia 17-1234-01); Immobiline™ DryStrips (pH 4-8L, 18cm; Amersham Pharmacia 17-1233-01)

Stains:

- Bromophenol blue (Aldrich 11,439-1); Coomassie® Brilliant Blue G250 (Fluka 27815); Coomassie® Brilliant Blue R250 (Serva, 17525)

2.1.4.2 Buffers

Sonication buffer:

- 10mM Tris/HCl pH 7.5; 5mM MgCl₂; cool down to 4°C

DNase I stock solution/ RNase A stock solution:

- 10mg/ml in H₂O; store in −20°C

Denaturation buffer:

- 1g sodium dodecyl sulphate; 0.23g 1,4-dithio-DL-threitol; add H₂O to 10ml final volume

Equilibration buffer:

- 0.4g 3-[(3-cholamidopropyl)dimethyl-ammonio]-1-propanesulfonate; 5.4g urea; 500µl Pharmalyte® pl 8-10.5; 0.1g 1,4-dithio-DL-threitol; add H₂O to 6.5ml final volume

Equilibration buffer:

- 0.4g 3-[(3-cholamidopropyl)dimethyl-ammonio]-1-propanesulfonate; 5.4g urea; 500µl Pharmalyte® pl 8-10.5; 0.1g 1,4-dithio-DL-threitol; add H₂O to 6.5ml final volume

Rehydration buffer:

- 12g urea; 0.5g 3-[(3-cholamidopropyl)dimethyl-ammonio]-1-propanesulfonate; 37.5mg 1,4-Dithio-DL-threitol; a trace of bromophenol blue and 0.5ml Resolyte® pH 4-8; add H₂O to 25ml final volume
CHAPTER 2 SEPARATING AND VISUALISING PROTEINS

Stock solution A (acrylamide 30%, NN'-methylenebisacrylamide 0.8%):

260g acrylamide; 7.0g NN'-methylenebisacrylamide; add H2O to 870ml final volume

Stock solution B (Running Gel buffer):

Trizma Base 1.5M (110g Tris), adjust pH to 8.8 with concentrated HCl; add H2O to 600ml final volume

Stock solution C (10% SDS):

2.5g sodium dodecyl sulphate (BDH 442442F); add H2O to 25ml final volume

Stock solution D (APS 10%):

2g ammonium peroxodisulfate; add H2O to 20ml final volume

Electrode buffer:

111g Tris; 573.6g glycine; 38.2g sodium dodecyl sulphate (Fluka 71729); add H2O to 38.2l final volume

Displacing solution:

50% glycerol; add H2O to 200ml final volume and add a trace of bromophenol blue

Equilibration solution

20ml Tris 500mM pH 6.8 stock solution; 72g urea; 60ml glycerol; 4g sodium dodecyl sulphate (BDH 442442F); add H2O to 100ml final volume

2.1.4.3 Cell extract preparation for Bradyrhizobium japonicum

Cells were grown to an A600 = 0.8 and harvested in a GSA rotor at 6000 rpm for 15 minutes. The pellet was washed twice in 0.9M NaCl and finally resuspended in sonication buffer. The pellet was stored at -20°C. For a typical 25cm x 25cm gel, 100mg pellet (~7-8mg protein) was taken and resuspended in 200μl sonication buffer. Care was taken to keep the sample concentration as high as possible to increase the final loading capacity. The cell suspension was sonicated on ice for ten minutes with a micro sonication tip (pulsed sonication, output control 3.5, duty cycle 10).

Adding 50μg DNase I and 10μg RNase A per sample digested the nucleic acid content. DNA and RNA were digested by incubation for 30 minutes at 37°C on a shaker. The sample was checked afterwards if the digestion was complete. An incomplete digested sample has a higher viscosity than complete digested samples. When DNA/RNA was still present further DNase and RNase was added.

The proteins were precipitated by ethanol precipitation. Ethanol was added to the solution to a final percentage of 75-80%. The solution was kept at -20°C for at least 30 minutes. The proteins were precipitated at 4°C at maximal speed in an Eppendorf centrifuge 5415 C for 20 minutes. The supernatant was removed and the pellet air-dried. Finally, the pellet was resuspended in 100μl sonication buffer.
The sample was denatured by adding solid urea until saturation was reached (two phases will appear and the solution is clearer) and incubated for 30 minutes at room temperature. Afterwards, denaturation buffer and equilibration buffer was added at ~1/10 of sample volume and incubated for one hour at RT on a shaker.

The solution was centrifuge twice for 30 minutes at maximum speed in an Eppendorf centrifuge and the pellet discarded. 3-6μl Resolyte® pH 3.5 - 10 were added to the supernatant using a sterile syringe.

2.1.4.4 The 1st Dimension - Immobiline rehydration and loading of the sample

250μl of sample were mixed with 350μl of rehydration buffer. The sample should not be heated or frozen at that stage and immediately loaded on the Immobiline™ DryStrips drop by drop. The strips were covered with Silicone oil DC 200 (1-2ml, no bubbles) in a homemade rehydration chamber and let stand over night (minimum 6 hours) at room temperature. The silicone oil stops the urea from crystallising on the Immobiline™ DryStrips. Crystallisation causes an inhomogeneous current in the first dimension run, leading to incomplete isoelectric focusing.

The first dimension running device (e.g. Pharmacia Multiphore II) was cooled down to 8°C. Care has to be taken not to cool down below 8°C, because the concentrated urea solution would crystallise.

The electrode wicks were cut and wetted with H₂O. To ensure thermal conduction, Silicon Oil DC 200 was poured between the Multiphore plate and the Immobiline tray as well as between tray and plastic foil. Bubbles should be avoided, because they slow down the heat transfer. The strips were quickly transferred onto the tray, aligned, the electrode wicks and electrodes carefully put on the strips, finally covered by silicon oil and the first dimension run started.

Power supply settings:

- 0-300 V, 50mA in 2 hours
- 300-3500 V, 50mA in 4 hours
- 3500-3500V; 50mA in 2 hours
- 3500-3500V, 50mA in 100 hours ---- total V x h should be 70000

2.1.4.5 The 2nd dimension - slab gel casting and transfer on second dimension

Using the ISO-DALT apparatus (Hoefer), 20 gels can be run in parallel. For casting 22 large gels, 870ml of stock solution A, 600ml of stock solution B, 25ml stock solution C, 20ml stock
solution D, 40 liters of electrode buffer and 200 ml of displacing solution was prepared. Due to the design of the chamber, the gels have to polymerise from the top down; hence, a gradient of APS and TEMED is made. The two solutions are named LIGHT and HEAVY although the concentration of acrylamide is the same, they differ in the amount of polymerisation inducing agent and glycerol. The electrode buffer was cooled down to 8°C before starting the run.

<table>
<thead>
<tr>
<th></th>
<th>LIGHT</th>
<th>HEAVY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrylamide/Bis</td>
<td>430 ml</td>
<td>430 ml</td>
</tr>
<tr>
<td>1.5M Tris/HCl</td>
<td>275 ml</td>
<td>275 ml</td>
</tr>
<tr>
<td>H2O</td>
<td>373 ml</td>
<td>290 ml</td>
</tr>
<tr>
<td>Glycerol</td>
<td>0 ml</td>
<td>88 ml</td>
</tr>
<tr>
<td>SDS 10%</td>
<td>11 ml</td>
<td>11 ml</td>
</tr>
<tr>
<td>APS 10%</td>
<td>11 ml</td>
<td>5.5 ml</td>
</tr>
<tr>
<td>TEMED</td>
<td>140 μl</td>
<td>48 μl</td>
</tr>
</tbody>
</table>

TEMED and APS were added to the heavy solution first, because the light solution tends to polymerise within 10 minutes, hence casting and overlay by a mixture of isopropanol/water must be completed quickly after the solutions were poured with the gradient maker. Depending on room temperature the gels polymerised in 2-3 hours.

After the first dimension run was completed, the electrodes were disconnected and the silicone oil was discarded. The strips were equilibrated on the tray. In a first equilibration step, the strips were reduced for 25 minutes in 100 ml equilibration solution with 2 g 1,4-Dithio-DL-threitol added. Subsequently the protein strips were alkylated for 5 minutes in 100 ml equilibration solution with 2.5 g iodoacetamide. For one tray, 100 ml equilibration solution was needed. The electrodes were taken away and the strips were subsequently transferred on top of the slab gel and sealed with agarose gel [50 ml of 1% agarose in 50 mM Tris/HCl pH 6.8, add a trace of bromophenol blue, heat to dissolve]. When the agarose gel became solid, the gels were placed in the tank and the overnight run was started at 420 mA total current.

When the run was finished, the gels were fixed for 3 hours, stained in Serva G250 for 2 hours and subsequently destained over night.
2.2 Visualising Proteins

To visualise proteins in gels, especially 2D gels, several methods are available. The method of choice always depends on the next step in the overall experimental strategy. If the separated proteins are subsequently analysed by mass spectrometry, silver staining disturbs the ionisation process (Gharahdaghi et al., 1999). Edman analysis requires blotting on PVDF, so staining methods have to be applied, which do not require fixation of the proteins in the gel. If the proteins should be quantitated, a procedure has to be chosen where the stain intensity is linear over a certain range and binding to proteins is stoichiometric.

The most widely used techniques to detect protein spots on two-dimensional gels are Coomassie® Blue and silver staining. The use of silver staining for quantitation is limited, because the relationship between silver staining density and protein concentration is characteristic for each protein (Merril et al., 1984). The amino acid composition and post-translational modifications of a protein spot determine the final quantity of silver, which is observed in the gel. The intensity of silver staining is linear over 40-50-fold range in concentration from 0.04ng/mm² to 2ng/mm² (Dunn, 1987). This corresponds well with the less sensitive Coomassie® Blue 20-fold staining range (10-200ng). It must be noted that protein-specific staining curves have also been observed for Coomassie® Blue (Tal et al., 1985), but the greater compatibility with the subsequent detection and identification tools has made it the staining procedure of choice. This became feasible especially since the loading capacity by rehydration allowed a significant improvement in the amount of protein loaded on a single gel. The methods are far from being optimal, so methods to overcome the problem of staining variability from gel to gel have to be developed.

To perform a comparison using the 2D-gel analysis program as used in this work, all the gels reflecting a certain time point in a kinetic or physiological state are used to create a single master-image representative for these conditions. A super-master-image is the created for the entire time set using all the master images created before. The integrated optical density of the spots are then calculated as a percentage of the total density of the super-master gel and the values are processed using a statistical data analysis program. A kinetic analysis of heat shock protein induction can be found in chapter 6 (Proteome Analysis of Heat Shock Protein Expression in *Bradyrhizobium japonicum*).
Other visualisation methods exist, such as radiolabelling specific isotopes and detection, the phosphor-imaging technology or \[^{35}\text{S}]\text{ thiourea silver staining}\ (Wallace and Saluz, 1992). Radiolabelling is highly sensitive, but has the disadvantage that the amount of label incorporated by each protein is dependent on the amino acid composition. An even more serious drawback is that labelling has to be done in culture, which requires specialised facilities and an enormous amount of expensive radioactive material (and waste).

Fluorescent stains as SYPRO Orange and Red (Steinberg et al., 1996; Steinberg et al., 1996) have some advances and will evolve into the stains of choice in the future. The visualisation is fast, does not require fixation (so the gel can be blotted afterwards) and can be applied during the transfer from the first to the second dimension. However, the staining is not uniform for all proteins and depends on the amount of SDS bound to a single protein, which is dependent on the amino acid composition and post-translational modifications.

Taking into account the above mentioned difficulties, stoichiometric and linear protein staining remains still a challenge and must be solved, if protein quantitation and kinetic studies have to be carried out in a fast and reliable way.
2.3 Application of the Techniques

The 2D-PAGE approach is very useful and is able to visualise broad changes in protein pattern of organisms exposed to different conditions in their environment. It can be used to analyse complex regulatory mechanisms such as heat-shock, because the induction or repression of several genes in the wt organism or mutants can be displayed simultaneously. As an example, the regulation of heat shock proteins in *Bradyrhizobium japonicum* has been chosen, because of the complex regulatory network helping the nitrogen-fixing root-nodule symbiont of Soya bean to cope with heat stress. A short introduction to the regulatory network is given, and the effect of a *hrcA* mutant on CIRCE/HrcA controlled genes is demonstrated.

2.3.1 The biological problem

The survival of any organism essentially depends on the ability to adapt to changes in the environment. Eubacteria have evolved a vast repertoire of positive and negative mechanisms to co-ordinately regulate the synthesis of heat shock proteins, which assist in protein folding, assembly, disassembly, transport and degradation.

*B. japonicum* uses an interdependent complex regulatory network of positive and negative mechanisms to control the expression of heat shock genes after a temperature up-shift (see Narberhaus, 1999). The use of the alternative sigma factor RpoH, the *E. coli* $\sigma^{32}$-analog, allows a transient induction of the *rpoH* regulon, comprising *dnaK*, *hrcA*, *grpE* and *groE* operons. The response is feedback controlled by the availability of DnaK. The concept of this homeostatic control and the involvement of the protease FtsH are explained in chapter 3, where the preparation of antibodies against the three known RpoH is described.

Two different negative heat shock regulatory mechanisms are known in *B. japonicum*. The conserved element ROSE (*repression of heat shock gene expression*; Narberhaus et al., 1998) controls at least five heat shock operons, which encode small heat shock proteins, RpoH$_1$ and other proteins. RpoH$_1$ connects the negatively acting ROSE mechanism to the positive control by RpoH (see Figure 2-1). The expression of small heat shock proteins and an overall proteomic view of the heat shock phenomenon in *B. japonicum* as well as in other various *Bradyrhizobium* and *Rhizobium* species is described by a series of two papers in chapter 6.
The third regulatory mechanism, the CIRCE/HrcA system (Roberts et al., 1996; Zuber and Schumann, 1994) is connected to the positive control mechanism by RpoH (see Figure 2-1). CIRCE (controlling inverted repeat of chaperone expression) is a negative cis element, found upstream of chaperone genes in a great number of bacteria (Hecker et al., 1996). The inverted repeat serves as a HrcA binding site, which is a negative regulator of the groESL$_4$ and groESL$_5$ operon in *B. japonicum*. A GroEL-mediated feedback mechanism controls the CIRCE regulon (Babst et al., 1996). HrcA depends on GroEL to acquire an active conformation. In the active conformation, HrcA is able to bind and repress the CIRCE regulon. GroESL$_4$ itself under CIRCE control provides the cell with a basal level of GroEL under normal temperature. After heat shock, the GroEL chaperones bind to denatured proteins. HrcA is rendered inactive and the CIRCE-dependent operons are transcribed. After a heat shock, GroEL is releasing the refolded proteins, binds to HrcA, renders it active and shuts down chaperone expression. In *B. japonicum*, *hrcA* is under control of the RpoH system. RpoH$_2$ seems to be responsible for the basal transcription of HrcA, providing the cell with a sufficient amount at normal temperatures (Minder et al., 1999). During a heat shock, *hrcA* is probably transcribed in a RpoH$_1$ dependent way, providing the cell with a high quantity of HrcA.
To confirm this model, a hrcA mutant was constructed. In HrcA-depleted cells, the transcription of the CIRCE-associated chaperonin operons groESL4 and groESL5 was strongly elevated even at physiological conditions (Minder et al., 1999). By two-dimensional gel analysis, complementary data for the derepression of the groESL4 and groESL5 operons were obtained.

If one compares Figure 2-2 A with the 2D gel of the wild type under normal growth conditions shown in chapter 6, a clearly elevated level of GroEL proteins is detectable. Under heat-shock conditions even an increase in the expression of GroEL proteins compared to physiological temperatures can be seen. Despite of being under CIRCE/HrcA control, the groESL4 and groESL5 operons are clearly transcribed under normal conditions. A leaky transcription from CIRCE-regulated promoters is also known for Bacillus subtilis or Clostridium acetobutylicum (Narberhaus and Bahl, 1992; Schmidt et al., 1992). It thus appears that the CIRCE/HrcA system is not a tight on/off switch, thereby providing the cell with major chaperones required for folding during normal growth. The 2D-gel analysis provides additional proof for the repressor effect of HrcA, showing an impressive increase of GroEL proteins under normal and heat shock conditions.
2.4 Discussion

Most physiological changes in a cell are associated with quantitative variations in the amount of gene products. Very seldom is an on/off situation found where a protein spot suddenly appears or disappears. Biological changes are gradual, so there is a need to quantitate the appearance or disappearance of a protein. Unfortunately, studies of gene expression at the transcript mRNA level alone do not take into account the occurrence of gene regulation at the post-transcriptional level. Anderson et al. (Anderson and Seilhamer, 1997) and Haynes et al. (Haynes et al., 1998) showed a general lack of correlation between the levels of mRNA and the resulting protein levels in liver and yeast cells, respectively. This inconsistency together with the occurrence of post-translational processing and modifications makes it necessary to examine the proteome to complete DNA sequence information and mRNA transcription patterns, if physiological changes in a cell are to be monitored. To separate and visualise a broad range of proteins, 2D PAGE and Coomassie® Blue/Silver Staining is the method of choice in many laboratories to assess protein purity and to separate proteins for analysis.

Recent developments in this field aim at the replacement of 2D electrophoresis. The Hochstrasser laboratory (Bienvenut et al., 1998) in Geneva is testing a 2D system, in which after a completion of the 2D run, the gels are blotted onto a double membrane. The first layer is a membrane with immobilised Trypsin, the second a standard PVDF membrane. In the first layer, the separated protein spot is digested into tryptic peptides, which are immobilised on the second layer. The second layer can be analysed directly by MALDI-TOF MS resulting in a peptide mass fingerprint (PMF, see chapter 5). This method is considered to be fast, but the informational content depends strongly on the quality of database used for PMF and the whole blot area (up to 25x25 cm) has to be scanned. If one presumes that every spot with an average area of 4mm² in the relevant gel area of 20x20 cm has to be scanned, 10000 measurements have to be done, which is very time (CPU and real) consuming to analyse. An average MALDI analysis in the full-automated mode takes about 2 minutes. To measure 10000 spots, a continuous run of 334 hours (~14 days) is needed to reveal in principle the whole proteome. The analysis is limited to PMF, because desorption of peptides immobilised onto PVDF membranes is only possible to achieve with large losses. Complex computer algorithms are needed to combine the data information into a virtual 2D-proteome map.
A similar approach is chosen in the Proteome research lab of Hoffmann La Roche in Basel (H. Langen, communicated by P. James). The 2D gels are cut into a grid of thousand little cubes, in-gel digested by Trypsin (see Chapter 4) and subsequently analysed by MALDI-TOF or MS/MS. This method provides the chance to obtain PMF, peptide fragment fingerprinting (PFF; see chapter 5) and de novo sequencing information (see chapter 4). Specialised gel cutters cut a 25x25 cm 2D gel in ~6500 gel pieces, in-gel digestion is carried out by robots and the MS/MS analysis is done in a full-automated mode. The data obtained are subsequently fused into a virtual 2D map by a computer program. This method is very time consuming and the application is still in the test phase. But with further improvements in the speed and sensitivity of the PMF, PFF and sequencing methods, this approach will be able to provide data complementary to the ongoing DNA sequencing projects.

An even more radical approach drops the separation step by 2D and replaces it with a capillary electrophoresis system, directly linked to an Fourier Transform (FT) MS (Pacific Northwest National Laboratory; S. Martinovic, personal communication). In the first dimension, capillary isoelectric focusing (CIEF) separates the protein mixture, which is subsequently online measured by FTMS to obtain the mass/second dimension. Therefore, a virtual 2D map (2D-display) is generated. This approach sounds promising, but the separation by capillary electrophoresis is restricted on soluble proteins and the separation is not very reproducible in the molecular range above 30kDa. The detection by Fourier transform ion cyclotron resonance mass spectrometers achieves high mass ranges, high resolution and simultaneous recording of all the different mass ions in the spectrum. The mass analysis data can be completed with low energy CID experiments, which also give sequence information of the virtual protein spot.
2.5 Reference


3 Detecting proteins

There are a lot of well-established techniques available for protein detection. The UV absorbance of the peptide bond at 215nm and of aromatic amino acid side chains at 280nm allows protein detection and quantification with a spectrophotometer. Other techniques such as IR-, CD- or NMR spectroscopy allows one to obtain additional information about the secondary or even tertiary structure of the protein. Probably the most commonly used separation tools in protein chemistry laboratories are capillary zone electrophoresis (CZE) and high-pressure liquid chromatography (HPLC). High-resolution columns and specialised packing material can be combined with a wide variety of detection methods such as UV, NMR or MS and allow the separation and detection of almost all types of biological material. The narrow (<300μm) and relatively long (5-50cm) capillary HPLC columns are packed with a noncompressible matrix coated with a thin layer of the stationary phase. The mobile phase is forced through the tightly packed column with pressures up to 5000psi. Peptide mapping by HPLC has been used quite often for protein detection, usually by comparison with a standard protein run at the same time. A fast way to detect proteins or peptides in solution is by mass spectrometry. In this chapter, only the principles of mass spectrometry will be explained and the use of MS as a simple detection / control method to monitor a complex peptide synthesis. In chapter 4 the use of mass spectrometry as an identification tool will be described and compared to traditional Edman sequencing. An alternative method for the detection of a protein in a complex mixture is the use of specific antibodies. The use of immunological methods is fast, sensitive and depending on the specificity, relatively reliable. Since antibodies can be prepared relatively fast and easily, they are widely used in many important analytical and preparative biochemical procedures. Immunodetection, mass spectrometry and microseparation methods all have an important role to play in proteome research. All can be used in automated procedures for protein identification. The following chapter will give a short introduction to mass spectrometry and describe the chemical synthesis of peptides for use in antibody generation. The different strategies for peptide synthesis and the application of HPLC and mass spectrometry as analytical and monitoring techniques will also be illustrated by the synthesis of three peptides from RpoH1, RpoH2 and RpoH3. These were used to generate and purify antibodies against the three peptides of σ32-factors. An example of the utility of these antibodies will then be shown to demonstrate the power of this approach. By in vivo stability experiments of the σ32 - type heat-shock factors it could be shown, that RpoH1 is FisH-resistant in contrast to RpoH2 and E. coli RpoH (Urech, C., Koby, S., Oppenheim, A., Müñchbach, M., Hennecke, H. and Narberhaus, F. σ32-type transcription factors that are resistant to degradation by the FisH protease in vivo. Submitted to J Biol Chem).
3.1 Mass Spectrometry

Mass Spectrometry is mostly used to generate two types of information: the molecular mass of an intact compound and structural information from the fragmentation pattern.

The generation of a mass spectrum involves three steps: 1. Generation of ions 2. Separation of ions according to their mass/charge ratio 3. Detection of ions.

3.1.1 Sample Ionisation

The main problem in analysing peptides and proteins by mass spectrometry is to bring these very labile and involatile large molecules into the gas phase prior to analysis in the high vacuum region in the mass spectrometer. Originally, the peptides and proteins had to broken down into di- and tripeptides and then extensively derivatised in order to make them sufficiently volatile. The major breakthrough came with the introduction of Fast-Atom-Bombardment (FAB) in 1981 by M. Barber (Barber et al., 1981). The peptides (up to mass 3,000) were introduced to the gas phase by sputtering from a relatively involatile matrix such as glycerol. The target matrix containing the peptides is bombarded with high velocity neutral argon atoms. The kinetic energy of the impacting ions led to sputtering of matrix into the vacuum and subsequent desolvation of the peptide. The method is moderately gentle and resulted in only a small amount of peptide fragmentation.

The two most popular techniques used today are electrospray ionisation (ESI; (Dole et al., 1968) and matrix assisted laser desorption and ionisation (MALDI, (Karas and Hillenkamp, 1988).

3.1.1.1 Electrospray Ionisation

ESI generates ions by spraying a liquid at atmospheric pressure in the presence of an electric field. The liquid is either acidified if positive ions are to be detected or made alkaline for negative ions. Electrospray is a two-stage process; charged droplets are created by evaporation from a liquid surface in the presence of a strong electric field, followed by evaporation of the charged droplets to yield charged sample molecules. As the droplets are formed, they carry a large number of analyte molecules in the presence of a vast excess of either H⁺ or OH⁻ depending on the pH. Half of the droplets are lost since they carry a net negative charge, the rest are accelerated toward the inlet of the MS. The solvent rapidly evaporates and the surface area
of the droplet decreases causing the net charge density to increase. At a critical value the charge repulsion force overcomes the surface tension and the droplet fragments into a series of very small droplets. The process is then repeated over and over until finally a solvated peptide or protein enters the MS. Here it passes down a heated capillary and emerges at the entrance of the analysis system as a 'naked charged molecule'.

3.1.1.2 Matrix assisted laser desorption and ionisation

Laser beams had been used fairly often for the generation of ions from the solid phase. It was not until Karas and Hillenkamp (Karas and Hillenkamp, 1988) introduced a sample preparation method in which the protein or peptide is embedded in a large excess of a laser light absorbing matrix (MALDI). This made the ionisation 'soft enough' to allow desorption and ionisation of large molecules without fragmentation. Ions are generated in MALDI by the absorption of pulses of laser energy from an UV or Infrared laser by a matrix. The compound or mixture thereof which forms the matrix, is designed to absorb at the wavelength produced by the laser. For UV lasers such as the most commonly used nitrogen (334nm) and Nd:YAG (266nm) lasers, compounds like α-cyano-4-hydroxycinnamic acid (α-CN), 2,5-dihydroxybenzoic acid (DHB) or 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid) are used as the matrix. For IR lasers, organic acids such as fumaric and succinic acid are the matrix of choice. The mechanism of ion formation and desorption is still not well described but essentially the laser energy absorbed by the matrix is passed by photon transmission throughout the crystalline matrix causing small regions to disintegrate and explode into the gas phase. Since the matrix is acidified, most of the proteins and peptides are protonated and are accelerated away by the high voltage field.

3.1.2 Mass Analysers

Mass analysers separate ions by making use of appropriate electric fields, sometimes in combination with magnetic fields. There are main three ways of separating ions commonly used in mass spectrometry.

3.1.2.1 Time-of-flight mass spectrometry

Time-of-flight mass spectrometry (TOF-MS) is based on the measurement of the time taken for an ion after generation by a laser pulse to reach the detector. After the ions are formed, they are accelerated by a high potential (ca. 30kV) towards the detector. Since the acceleration is inversely proportional to the mass, the time taken to reach the detector can be translated into a
mass range. The time range can be calibrated using known molecular weight standards and used to determine exact masses. The ions formed by the initial laser pulse show a wide spectrum of velocities, which leads to a spreading of the mass envelope. The introduction of delayed extraction allows ions to 'cool' down, narrowing the energy spread (Brown and Lennon, 1995). This together with the use of a reflectron instrument, which gives a longer flight path and hence better temporal resolution, allows very high sensitivity and accurate mass measurements.

3.1.2.2 Quadrupole mass spectrometry

Quadrupole mass spectrometry (Q-MS) was developed by Paul in the late 1950's (Paul, 1990). A quadrupole electric field is created by electric potentials applied to four parallel rods with an elliptical central cross section. Such fields can be used to filter ions according to their m/z ratio, but can be used to trap ions. A Q-MS can be made to act as an ion selective filter by scanning a dc Voltage $U$ across a pair of diametrically opposed rods in the presence of an Rf voltage $V_0 \cos (\omega t)$ which is placed across the other pair of rods. At a particular field strength, ions having a small range of m/z values have stable paths through the filter and other ions are not transmitted. They either collide with the rods or are pumped away into the vacuum system.

3.1.2.3 Ion trap mass spectrometer

A third type of a mass analyser is the ion trap mass spectrometer. It can be considered as a 3D analogue of the quadrupole mass filter. The trapped ions are analysed using a mass-selective instability scan in which the amplitude of the Rf voltage applied to the trapping ring is scanned, causing ions of increasing m/z to adopt unstable trajectories and to be ejected from the trap, where they are detected.

3.1.3 Detectors

After the ions are generated from the sample in the ion source and separated according to their mass-to-charge ratio in the mass analyser, each type of ion strikes the detector and produces a signal proportional to it’s relative abundance. A plot of the relative abundance of each ion against m/z yields a mass spectrum. Detectors for mass spectrometry fall into two categories: direct measurement detectors that detect the charge arriving at the detector such as Faraday cups, and secondary electron multiplier detectors, which detect the electrons ejected by the impact of the arriving ion after cascade amplification.
3.1.4 MALDI-TOF-MS

Most often, the MALD ionisation technique is used in time-of-flight mass analysers equipped with a micro-channel plate detector. The sample is co-crystallised with UV absorbing matrix. A laser impulse is used to desorb the sample from the matrix and to start timing the arrival of ions at the detector. The ions are accelerated by large potential differences (30kV) and pass down the evacuated tube to an ion mirror (reflectron). The mirror reflects the ions in a mass dependent manner to the detector. The mass is determined by measuring the time taken for the ion to travel from the target to the detector.

Figure 3-1: schematic diagram of reflectron MALDI-TOF-MS
3.2 Immunodetection

Most of the antibodies that are useful in biochemistry are of the IgG type. These are Y-shaped molecules, each with two antigen-binding sites. This multivalence can be used to create giant insoluble networks in precipitin reaction. In older analytical techniques as immunodiffusion or immunoelectrophoresis, the insoluble nature of this complex is used to characterise mutant organisms or quantitate antigens. Much more widely used to quantitate antigen-antibody reactions is the enzyme-linked immunosorbent assay (ELISA) or the radioimmunoassay (RIA) which forms the basis of many clinical diagnosis tests. Although these methods have a wide variety of variations, the principle is to assay for bound antibody simply by monitoring the activity of the conjugated enzyme or the radioactivity bound to the antibody. Western blotting is used to detect, in a complex mixture of proteins or fragments of proteins, those that react with the same antibody. The main drawback of this last method is the difficulty to align the electroblot with the gel, especially when large gels are blotted. Western blotting can be used to monitor the expression or degradation of a protein under different conditions in a cell or to follow the posttranslational cleavage of proteins in their maturation process. Because of their high specificity of protein binding, antibodies can also be used to purify proteins (immunoaffinity chromatography). They even can be used to visualise intracellular organisation and help to localise particular antigens in cytological preparations (immunocytochemistry).

But to apply all these techniques the protein or peptide of interest has to be obtained and the antibody recognising the protein or peptide of interest has to be generated. There are three major ways to obtain peptides or proteins: (1) the purification out of biological material (2) genetecno logical production or (3) peptide synthesis.

The aim of peptide synthesis is to produce fragments of proteins or whole proteins by organic chemical synthesis. The length of peptides, which can be obtained by chemical synthesis, is limited. However, unnatural or modified amino acids can be incorporated in the synthesised peptide using this powerful method. Moreover, it is possible to obtain pure and homogeneous material, which is of great importance for pharmaceutical products.
3.2.1 Solid phase peptide synthesis - the Merrifield technique

Because of the complexity of proteins classical methods of solution based peptide chemistry are inappropriate to synthesise peptides longer than five amino acids. One major problem is the difficulty to purify the product after each synthesis step, because the chemical behaviour of the peptide is significantly changed after a new amino acid is attached to the growing peptide chain. These limitations were overcome by the introduction of solid phase peptide synthesis by Bruce Merrifield in 1962 (Merrifield, 1967; Merrifield and Stewart, 1965).

The principle of solid phase peptide synthesis (SPPS) is simple. It is based on sequential addition of side-chain protected α-amino acids to an insoluble and stable polymeric support. The growing peptide chain remains attached to this particle throughout all the synthetic steps and is separated from soluble reagents and solvents by simple filtration and washing. Finally, the desired product is detached from the solid support. The purification and characterisation is carried out in free solution. There are a number of immediate consequences. The activation/addition/cleavage and washing steps are quick and simple. There is an enormous labour advantage over the corresponding operations in solution chemistry and it can be automated. But there are some inbuilt disadvantages. Pre-eminent amongst them is the need for very high reaction efficiencies. Using large excesses of reagents and feedback monitoring can reduce this problem. Another drawback lies in the solid phase nature of the system, which renders inappropriate many of the analytical procedures, especially the spectroscopic techniques. Not uncommonly, solid phase peptide synthesis is carried out completely blind until the final purification of the potential synthesised peptide.

3.2.2 Fmoc and Boc chemistry in SPPS

A polypeptide is synthesised starting from the C-terminus in direction to the N-terminus, opposite to the biological way. This fact implies, that each amino acid to be attached to the growing C-terminal polypeptide chain, must be N-terminal blocked and all side-chains protected. Otherwise amino acids could also react between each other and the reaction would not be restricted to the only free N-terminus bound to the growing peptide chain. After coupling of the free amino acid and formation of a new peptide bond, the new N-terminus must be deblocked that the next peptide bond can be formed, without the removal of the side-chain protection. So each cycle of amino acid addition requires a blocking and a deblocking step.
There are two major methods of N-α-protection using the acid-labile Boc group (t-butyloxycarbonyl) or base-labile Fmoc-group (9-fluorenylmethoxycarbonyl). After removal of this N-α-protection group, the next protected amino acid is added using a coupling reagent or pre-activated protected amino acid derivatives. During the entire peptide synthesis the peptide is attached to the resin via a linker through its C-terminus and may be cleaved to yield a peptide acid or amide, depending on the linkage agent used. Side-chain protecting groups are often chosen to be cleaved simultaneously with detachment of the peptide from the resin. Cleavage of the Boc protecting group is achieved by trifluoroacetic acid (TFA) and the Fmoc protecting group by piperidine. Final cleavage of the peptidyl resin and side-chain deprotection requires harsh conditions, using hydrogen fluoride (HF) or trifluoromethanesulfonic acid (TFMSA) with Boc chemistry or TFA in Fmoc chemistry. dichlormethane (DCM) and N,N-dimethylformamide (DMF) are the primary solvents used for peptide deprotection, coupling and washing.

3.2.3 Protecting groups and cleavage strategies in Fmoc chemistry

In order to achieve milder reaction conditions the types of protecting group and resin-linkage systems used needed to be rethought. The introduction of 9-fluorenylmethoxycarbonyl (Fmoc) as a base-labile protection group for α-amino protection by Carpino and Han proved suitable (reviewed in (Albericio and Carpino, 1997). This allows the rapid cleavage by solutions of secondary bases such as piperidine in DMF (see Figure 3-2). The conditions used left even
Detecting Proteins Chapter 3

particularly sensitive t-butyl derivatives entirely unaffected. Thus, a complete selectivity between α-amino and side-chain deprotection was obtained.

Not all amino acids require side-chain protection. Those, which do, can undergo irreversible side reactions, either during the synthesis or during cleavage, if the protecting group is not compatible or if an inappropriate cleavage method is used. The choice of side-chain protection is not only dependent on the chemistry used, but also on the coupling and cleavage methods employed, solubility of the derivative and the sequence of peptide being synthesised. Because of the huge number of amino acid derivatives available, it would be unsuitable to discuss them in this context. Excellent reviews have been written and almost all information needed to choose the right protection strategy is provided in a paper by Stewart and Young (Stewart and Young, 1984). The overall strategy in Fmoc chemistry is the use of acid-labile protection groups, which can be removed in parallel with the Fmoc resin cleavage.

3.2.4 Coupling methods

There are several basic types of coupling techniques for step-wise introduction of N-α-protected amino acids in SPPS. Carbodiimides were the first described (Geiger et al., 1971). In situ activating reagents are now the most commonly accepted because of their ease of use and fast reactivity (Fields and Noble, 1990). Most are based on phosphonium or uronium salts, which, in the presence of a tertiary base can smoothly convert, protected amino acids to activated species. The most commonly deployed are PyBOP® (1H-Benzotriazole-1-yl-oxy-tris-pyrrolidinophosphonium hexafluorophosphosphate) TBTU (2-(1H-Benzotriazole-1-yl)-1,1,3,3-tetramethyl uronium tetrafluoroborate) and HBTU (2-(1H-Benzotriazole-1-yl)-1,1,3,3-tetramethyl uronium hexafluorophosphate) and they generate HOBt (N-Hydroxybenzotriazole) esters (Figure 3-3).

A crucial, often disregarded step on a solid support is the solvation of the peptide resin. The linear, growing peptide chains have to be accessible at each step otherwise there is a reduction in the overall yield which can be attributed to poor solvation of the growing chain and intermolecular peptide chain aggregates. Enhanced efficiencies can be seen upon the addition of polar solvents such as DMF and N-methypyrrolidone (NMP), but the solvents must be used in highest purity available and it is even recommended to distil the solvents prior to use.
3.2.5 Resin tests

It is often advisable to measure coupling efficiency by assessing the presence of free and unreacted amino groups. The most widely used qualitative test for the presence or absence of free amino groups (deprotection/coupling) was devised by Kaiser (Kaiser et al., 1970). This test is simple and fast, however it should be noted that some deprotected amino acids do not show the expected dark blue colour typical of free primary amino groups (Fontenot et al., 1991); e.g. serine, asparagine and aspartic acid). Proline, being a secondary amino acid, does not yield a positive reaction. Occasionally false negative tests are observed, particularly with strongly aggregated sequences, which shows that one has to be very careful when assessing the extent of deprotection or coupling on solid phase. For proline, the chloranil test is recommended (Vojkovsky, 1995).

3.2.6 Fmoc resin cleavage and deprotection

Having synthesised a protected peptide one has to simultaneously detach and deprotect the peptide without oxidising or altering it. In Fmoc SPPS, this step is normally done by adding TFA to the peptidyl resin. During this process, highly reactive cationic species are generated from the protecting groups (King et al., 1990). These can react and modify those residues, which contain nucleophilic functional groups: tryptophan, methionine, tyrosine and cysteine. To prevent this, scavengers (nucleophilic reagents) are added to TFA to quench these ions. A number of cleavage mixtures are known in the literature, but the most widely used is Reagent K (TFA/water/phenol/thioanisol/ethanedithiole (EDT) [82.5:5:5:5:2.5]. To avoid toxic EDT a mixture of 95% TFA 2.5% Triisopropylsilane and 2.5% water can be used.
Figure 3-4: The peptide synthesis strategy; details are given in text.
3.2.7 Post-cleavage work-up

Before precipitation by ether or t-butyl-methyl ether, TFA is removed using a rotary evaporator, equipped with a CO₂/acetone cold finger, oil pump and acid trap. Filtering of the precipitated peptide through hardened filter paper in a Hirsch funnel under light vacuum removes the scavengers and protecting groups. An often recognised modification is the acid-catalysed oxidation of methionine to sulfoxide during the TFA cleavage- and deprotection-step. The best way to re-reduce methionine is described by Beck (Beck et al., 1994). The crude peptide is then subjected to semi-preparative RP-HPLC to desalt and separate the full-length synthesised peptide from all incomplete synthesis products. Usually a packing of C₈ or C₁₈ is used and a water/MeCN gradient with acidic ion pairing reagents like TFA is used. The monitoring is performed using a wavelength of 215 nm and if the peptide contains aromatic residues, monitoring of 280 nm can also be performed. The peaks are analysed for the correct mass with MALDI-TOF-MS. The peptide is subsequently lyophilised and is stored under argon frozen at -20°C. It is recommended to resuspend the peptides in water and lyophilise again to remove all residual acid, because acid can slowly react with unprotected side-chains, N-terminal glutamine or even hydrolyse acid-labile peptide bonds as Asp-Pro.

3.2.8 Antibody generation

Low molecular weight molecules are often not sufficiently immunogenic to elicit an immune response. Synthetic peptides can be made immunogenic by conjugation to a suitable carrier. A wide range of polymeric carriers and conjugation methods are available to prepare immunogens from small non-immunogenic haptens. These include carrier proteins as bovine serum albumine (BSA), keyhole limpet hemocyanin (KLH), thyreogloblin (TG) or synthetic carriers such as multiple antigenic peptides (MAPS). KLH, a large molecular weight protein (MW 2-3 x 10⁶) isolated from the mollusc Megathura crenulata, is often used due to its high immunogenic properties and the large number of lysine residues available for modification. Haptens can be coupled to carriers by the use of bifunctional cross-linkers such as glutaraldehyde (GA) or heterobifunctional cross-linkers such as N-hydroxsuccinimidyl esters of maleimido-alkyl-carboxylate derivatives. The conjugated covalently-linked peptide-protein conjugate is then ready to be used in an immunisation reaction.
There are two main techniques available, to generate antibodies. With the in vivo approach, an immunogenic hapten is injected into an animal and after an immune reaction, antibodies against the injected hapten can be obtained. With phage display, tailor-made antibodies may be synthesised and selected to acquire the desired affinity of binding and specificity (Hoogenboom et al., 1998). This second, in vitro approach is very powerful and is becoming a viable alternative to the use of animals. The former, traditional approach requires the synthesis of an immunogenic peptide. Antibodies against RpoH₁, RpoH₂, RpoH₃ from Bradyrhizobium japonicum were generated in order to study the expression and stability of these σ₃₂-type transcription factors.

### 3.3.1 The synthesis of three peptides of RpoH₁, RpoH₂ and RpoH₃

Three peptides were chosen and the only methodological restriction given was that they should contain a sulfhydryl group to link the hapten to maleimide-activated KLH (keyhole limpet hemocyanin), a widely used carrier protein with high immunogenic properties.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Sigma-factor</th>
<th>Sequence</th>
<th>MW (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA 1</td>
<td>RpoH₁</td>
<td>MQTSHEVARS ASVAAAGAAC</td>
<td>1916</td>
</tr>
<tr>
<td>MA 2</td>
<td>RpoH₂</td>
<td>CSAVKGTIAR AEQAALBAAH</td>
<td>1995</td>
</tr>
<tr>
<td>MA 3</td>
<td>RpoH₃</td>
<td>CANGPPSSRA KEMKQGVARA</td>
<td>2056</td>
</tr>
</tbody>
</table>

*Table 3-1: Sequences and MW of the generated peptides of the three RpoH*

The Fmoc chemistry is the method of choice in our laboratory due to the less toxic chemicals needed. It allows one to avoid the potential alteration of sensitive peptide bonds as well as potential acid catalysed side reactions during the repetitive TFA acidolysis in Boc chemistry. In the cleavage of Boc groups, the total exposure to TFA may exceed 7 hours during the synthesis of a 20-residue peptide. For SPPS the rink amide resin from Novabiochem (4-(2',4'-Dimethoxyphenyl-Fmoc-aminomethyl)-phenoxy resin attached to 1% cross-linked aminomethyl-polystyrene support) with a loading of 0.47mMol/g was chosen. The resin has an ideal swelling behaviour in organic solvents as NMP or DMF and must be preswollen for at least 1 hour before the start of the synthesis. 0.2g of resin was weighted and filled in Teflon
synthesis vials on a Teflon peptide shaker equipped with a washing device. The protecting Fmoc-group was removed with 30% Piperidine in DMF and the resin was extensively washed (5x DMF, 2x Acetonitrile (MeCN)). See steps (1)-(3) in Figure 3-4. In the case of resin-bound Cys(Trt), the stronger base 1,8-diazobicyclo[5.4.0]undec-7-ene (DBU) was used, because it reduces the enantiomerisation of Cys(Trt) (Wade et al., 1991). DBU was also used for Fmoc-deprotection with peptides longer than 15 amino acids to overcome incomplete deprotection by piperidine which occurs with longer peptide chains showing increasing secondary structure (Fontenot et al., 1991). A positive Kaiser Test (ninhydrin reaction) showed a successful deprotection of the resin. The test was performed by adding 15μl each of solution A (5g ninhydrin in 100ml ethanol), solution B (80g liquefied phenol in 20 ml ethanol) and solution C (2ml of 1mM aqueous solution of potassium cyanide in 98 ml pyridine) to a few resin beads. After heating to 120°C for 5 minutes a positive test is indicated by a blue resin colour. The C-terminus of 0.5mmol side-chain protected amino (2.5 times excess) was activated with 0.45mmol coupling mixture (1400μl of 4g TBTU and 30g HOBt in 37.5ml NMP/MeCN (2:1)) in sym-Collidine. The reaction was carried out for 15 minutes at room temperature. See step (4)-(7) in Figure 3-4. After extensive washing and a negative Kaiser Test, the unreacted free amide groups were capped by acetic anhydride in N-ethylisopropylamine (Hünig’s base). After the complete removal of the capping reagent a new cycle is started with the Fmoc-deprotection step described above. Before complete Fmoc resin cleavage, the peptide resins had to be thoroughly washed to remove excess of DMF, because it is non-volatile and its basic character can have inhibitory effects on TFA-acidolysis. It also absorbs strongly wavelength of 210 – 220nm, which interferes with the HPLC monitoring in the subsequent purification step. The beads were first extensively washed with dichloromethane to remove all the non-volatile solvents and were subsequent dried with ether. The peptide was cleaved from the resin by incubation for two hours at room temperature in 95% TFA, 2.5% Triisopropylsilane and 2.5% water. The crude product obtained was washed twice with TFA. A precipitation step with diethylether removed most of the released protecting groups and linkers. An analysis of the crude extract by MALDI-TOF-MS was carried out to determine the peptide containing fractions (data not shown). The Met-containing peptides RpoH1 and RpoH3 contained oxidised methionine (data not shown). The peptides were dissolved in 1ml TFA and set under argon atmosphere. 16μl of ethandithiol (EDT) and 13μl of trimethylbromsilan were added. After a reaction time of 15 minutes at room temperature, the peptides were precipitated with diethylether and isolated as described before. The crude peptides were analysed by reversed phase (RP)-HPLC on a capillary column (C18, 5μm, 300Å, 280x0.32mm) to detect the elution time of the major pep-
tide peak on an analytical HPLC system (ABI 140B solvent delivery system, ABI 1000S-diode array detector, Applied Biosystems). A preparative RP-HPLC run (C$_{18}$ preparative column; 250x21mm, Nucleosil 100–12, Macherey-Nagel; preparative reversed-phase HPLC system L-6220 Intelligent pump, L-4250 UV-VIS Detector, Merck AG) was done to purify the bulk amount of peptide. Fractions of interest were collected, three times lyophilised and subsequently analysed by an analytical RP-HPLC on a capillary column (C$_{18}$, 5μm, 300Å, 280x0.32mm). The masses of the synthesised peptides were confirmed by MALDI-TOF-MS (Voyager Elite MS, Perseeptive Biosystems).

The analytical HPLC runs and MALDI-TOF-MS are shown for RpoH$_1$, RpoH$_2$ and RpoH$_3$ in Figure 3-7, Figure 3-6 and Figure 3-5, respectively.
Figure 3-5: Purification and analysis of MA-1 (RpoH1). In A the analytical HPLC run (Macherey-Nagel, ET 250/2 Nucleosil 300-5 C18 250x2mm; ABI 140B solvent delivery system, ABI 1000S-diode array detector) of crude MA-1 is shown; the solvent B (80% Acetonitrile) gradient is indicated by the dotted line. In B the analytical HPLC run after the preparative purification of MA-1 is shown. The dotted line indicates the solvent B gradient. C shows the MALDI-TOF mass spectrum recorded in reflectron mode using a Voyager Elite MS (Perseptive Biosystems).
Figure 3-6: Purification and analysis of MA-2 (RpoH2). In A the analytical HPLC run (Macherey-Nagel, ET 250/2 Nucleosil 300-5 C18, 250x2mm; ABI 140B solvent delivery system, ABI 1090S-diode array detector) of crude MA-1 is shown; the solvent B (80% Acetonitrile) gradient is indicated by the dotted line. In B the analytical HPLC run after the preparative purification of MA-2 is shown. The dotted line indicates the solvent B gradient. C shows the MALDI-TOF mass spectrum recorded in reflectron mode using a Voyager Elite MS (Perseptive Biosystems).
Figure 3-7: Purification and analysis of MA – 3 (RpoH). In A the analytical HPLC run (Macherey-Nagel, ET 250/2 Nucleosil 300-5 C₁₈, 250x2mm; ABI 140B solvent delivery system, ABI 1000S-diode array detector) of crude MA-1 is shown; the solvent B (80% Acetonitrile) gradient is indicated by the dotted line. In B the analytical HPLC run after the preparative purification of MA-3 is shown. The dotted line indicates the solvent B gradient. C shows the MALDI-TOF mass spectrum recorded in reflectron mode using a Voyager Elite MS (Perseptive Biosystems).
Maleimide-activated KLH was chosen as a substrate to conjugate the obtained peptides. The use of maleimide as a cross-linking reagent results in a selective modification of the carrier protein at the ε-amino group of lysine side chains to form stable amide bonds. Activated KLH has reactive maleimide groups on its surface available for conjugation with a sulfhydryl-containing hapten to form stable thioether bonds. The reaction of the maleimido group and the sulfhydryl group proceeds rapidly and selectively under mild coupling conditions (pH 6.6) in 20mM sodium phosphate buffer, 0.23 M NaCl, 2mM EDTA and 80mM sucrose. 4mg of peptide was coupled to 2.5mg of maleimide-activated KLH (2000:1 molar ratio) and degassed, while stirring under a gentle nitrogen stream for two minutes. The reaction continued under stirring for two hours at room temperature. The obtained peptide-protein conjugates were now ready for immunisation. The immunisation of rabbits, serum extraction and purification steps were done by Dr. Lloyd Vaughan. ELISA showed a high reactivity of the obtained sera, when the RpoH peptides were displayed as antigen (data not shown). To reduce the cross-reactivity of the sera, a cross-purification was carried out. The peptides were covalently linked to cyano- gen bromide activated Sepharose. The sera were then cross-purified according to the following procedure: MA−1 serum was diluted 1:1 in PBS and absorbed over MA−2 Sepharose and MA−3 Sepharose columns linked in series, prior to binding the antibodies to MA−1 Sepharose. After washing, the antibodies bound to the MA−1 Sepharose were eluted with 0.1M Glycine, pH 2.5 and the pH was adjusted to 7 after elution with 1M Tris/HCl, pH 8.3. The elution was monitored with a spectrophotometer at 280nm and fractions of 0.5ml were collected (see Figure 3-10 on page 3-21). With a reference of IgG it was possible to calculate the protein content of the fractions. A 1% solution of IgG has an OD$_{280}$ of 6.8 in a 5mm path length cuvette. The fractions were stored at 4°C.
3.3.3 The application of the three antibodies $\alpha$-RpoH$_1$, $\alpha$-RpoH$_2$ and $\alpha$-RpoH$_3$

A first experiment demonstrates the selectivity of the RpoH antisera. Figure 3-8 indicates that $\alpha$-RpoH$_1$ specifically recognises *B. japonicum* RpoH$_1$. $\alpha$-RpoH$_2$ and $\alpha$-RpoH$_3$ are specific for RpoH$_2$ and RpoH$_3$, respectively. None of the purified antibodies recognises $\sigma^{32}$ of *E. coli*.

![Figure 3-8 Western Blot of cell extracts of E. coli C600/pRJ5007 (E. coli RpoH) — lane 1: E. coli C600/pRJ5040 (RpoH$_1$) — lane 2: E. coli C600/pRJ5002 (RpoH$_2$) — lane 3: E. coli C600/pRJ5007 (rpoH$_3$) — lane 4: in the left Western Blot anti-RpoH$_1$ is used, in the middle anti-RpoH$_2$ and on the right Blot anti-RpoH$_3$.

32kDa

The highly specific antisera allowed testing for the in vivo stability of RpoH$_2$ and RpoH$_3$. In experiments performed by Carmen Urech (Institute of Microbiology, ETHZ) *E. coli* C600 (wt), *ftsH* and *dnaK* were transformed with pRJ5002 and pRJ5040. The cultures were shifted to 42°C (non-permissive temperature for *E. coli* *ftsH*) for 15 minutes. Then chloramphenicol was added to stop de novo protein synthesis and the decay of RpoH proteins was examined by Western blot analysis. A comparison of the Western blots of *B. japonicum* RpoH$_2$ in *E. coli* wt and in *E. coli* *ftsH* revealed that RpoH$_2$ is unstable in the presence of FtsH. On the other hand, the degradation kinetics of RpoH$_3$ in wt and *ftsH*-mutant were identical, indicating that this sigma factor is not degraded by FtsH but by other proteases (see Figure 3-9). The degradation of *B. japonicum* RpoH$_2$ and *B. japonicum* RpoH$_3$ seems in both cases dependent on DnaKJ.
RpoH₁ is well recognised by α-E. coli σ-serum. It could be shown by Carmen Urich (Urech, 1999), that *B. japonicum* RpoH₁ is FtsH-resistant in contrast to *E. coli* RpoH and *B. japonicum* RpoH₂ (data not shown).

**Figure 3-9:** Stability of RpoH₂ and RpoH₃ in *E. coli*. For explanation see text. See numbers below indicating minutes after chloramphenicol addition.

**Figure 3-10:** Elution profiles of peptide-Sepharose columns. See text for detailed explanation.
3.4 Discussion

3.4.1 The detection problem

Peptide synthesis provides a fast and reliable way to generate a pure peptide of choice. This peptide can be used to generate antibodies or can be studied in other biological experiments such as NMR spectroscopy. Immunodetection is a sensitive and specific way to identify single protein spots from 1D or 2D PAGE, and it can detect as little as some femtomoles. Using this powerful method, Celis (Celis et al., 1995) identified components of the signal transduction cascade, including RAS, MEK2, Cip1 and Cdk2. However, the techniques are slow, taking into account the tedious preparation of an antibody. Even with an antibody at hands, only a few Western blots a day can be processed. Recently, a multiple immuno 2-D blotting (MI-2DB) method using several monoclonal antibodies was developed to simultaneously analyse oncogene expression and cell cycle checkpoints (Sanchez et al., 1997). Nine low abundance proteins could be detected simultaneously. MI-2DB is an excellent technique to display and quantitate expressed proteins that are simultaneously expressed and present in similar quantities.

The antibodies generated in this study were used to investigate the following biological problem.

3.4.2 The biological problem

In *B. japonicum* three different mechanism for heat-shock regulation are known. One of the mechanisms is dependent on three $\sigma^{32}$-type heat-shock factors. Not only the existence of an elaborate regulatory network, also the presence of three $\sigma^{32}$ factors makes the heat-shock regulation a very complex phenomenon. The expression of the individual *rpoH*-genes is controlled by different regulatory mechanisms (Narberhaus et al., 1997). The expression of *rpoH* is induced after heat-shock (43°C) and is under control of the so-called ROSE element (Narberhaus et al., 1998). $RpoH_2$ is constitutively transcribed from a $\sigma^{70}$ promoter. In addition, there is a $\sigma^5$ promoter, which controls the expression at very high temperatures (48°C). $RpoH_1$ and $RpoH_2$ show different promoter specificities (Narberhaus et al., 1998). The non-essential $RpoH_1$ controls the expression of GroESL after heat-shock. $RpoH_2$ seems to be essential for the expression of cellular proteins under physiological conditions and provides the cell with the DnaKJ under normal conditions.

In *E. coli*, the $\sigma^{32}$ heat-shock induction mechanism is extensively studied. The translation rate of *rpoH*-mRNA is higher and the $\sigma^{32}$ protein encoded by *rpoH* has a higher stability after heat-shock.
Detecting Proteins

Chapter 3

Under normal physiological states $\sigma^{32}$ is an extremely labile protein due to its degradation by FtsH (Herman et al., 1995; Straus et al., 1987; Tomoyasu et al., 1995). The involvement of FtsH, a so-called Charonine (protease-associated chaperonine; (Schumann, 1999)), led to the following proposed regulatory model (Hughes and Mathee, 1998). Under normal physiological states, DnaK functions as an anti-$\sigma^{32}$-factor forming a stable complex with $\sigma^{32}$ and DnaJ, resulting in inactive $\sigma^{32}$. This complex, together with GrpE, delivers $\sigma^{32}$ to the protease FtsH for degradation. Under heat-shock, the chaperones DnaK, DnaJ and GrpE release $\sigma^{32}$. The machinery binds preferably to denatured proteins accumulated under these conditions. $\sigma^{32}$ accumulated under these conditions is now free to bind to RNA polymerase core enzyme and is not degraded by FtsH, resulting in higher stability of $\sigma^{32}$ under heat-shock conditions and transcription of heat shock genes.

To study the role of the three $\sigma^{32}$-type heat-shock factors RpoH1, RpoH2 and RpoH3 in *B. japonicum*, three mutants were generated. Mutant H1 was depleted of rpoH1, mutant H3 of rpoH3 and the double mutant H1+3 of rpoH1 and rpoH3. It was not possible to generate viable rpoH2-defective mutants (Narberhaus et al., 1997). In preliminary experiments using the 2D-SDS-PAGE approach, it was not possible to detect any differences in the protein pattern of the three mutants, even after silver staining and computer (2D gels not shown). This might be explained by the presence of RpoH2 in all mutants.

To discriminate between the three heat-shock factors and to reveal their putative biological relevance in *B. japonicum*, it was decided to generate antibodies against peptides of the sigma factors. The generation of specific antibodies should facilitate comparison of the stability of individual sigma factors. In the diploma work performed by Carmen Urech (Institute of Microbiology, ETHZ). It was shown that the peptide antibodies specifically recognise the appropriate proteins in Western blot experiments. It was demonstrated, that RpoH2 is an unstable protein in *E. coli* and it is stabilised in cells depleted of FtsH. RpoH2 might thus be under the same regulatory mechanism as $\sigma^{32}$ in *E. coli*. The degradation seems also dependent on DnaKJ, because RpoH2 shows a higher stability in mutants depleted on DnaKJ. RpoH3 seems to be resistant to FtsH degradation, because no difference in stability could be detected in *ftsH* defective cells. The degradation of RpoH3 seems to be dependent on DnaKJ, because RpoH3 showed a higher stability in the dnaKJ-mutants. RpoH2 might be degraded by a so far unknown protease. It has been reported that *E. coli* $\sigma^{32}$ is not only a substrate of FtsH but also of other proteases such as HslVU or Lon (Kanemori et al., 1997; Kanemori et al., 1999). The stability of RpoH1 seems not to be dependent on FtsH. The latter result is in perfect agreement with previous studies obtained by F. Narberhaus (Narberhaus et al., 1996).
3.5 References


4 Identifying Proteins

In Proteome projects, which aim to identify and assign function to all proteins expressed by an organism at a certain time, the identification step is central. In the simplest sense, the identification assigns a name or database accession code to a protein or spot on a gel. A link between a gel spot, the amino acid sequence of a protein and the DNA sequence of a gene is obtained, thus linking genome and proteome. Identification is also the first step towards studies on protein function as well as co- and post-translational modifications, thus providing a link between proteome and metabolome.

The proteins expressed by an organism are separated and visualised on 2D gels as described in chapter 2. The whole protein gel can be screened by the use of antibodies, which can be generated against almost all haptens as described in chapter 3. To identify a protein, its primary and secondary attributes have to be revealed. A primary attribute is generated from the intact protein; a secondary is generated from fragments of the whole molecule. In the following chapter, the identification of proteins using Edman degradation chemistry (primary attribute), chemical tag sequencing and de novo sequencing by mass spectrometry (secondary attributes) will be explored. Chapter 5 will deal with characterisation programs to interpret data obtained by peptide mass fingerprinting and peptide fragment fingerprinting. To demonstrate the practicability of MS/MS sequencing, the analysis of a two-dimensional gel spot not present in the regR mutant of B. japonicum is shown. The technique of selective ion labelling simplifies the interpretation of MS/MS spectra to a great extent. Ion labelling by nicotinic acid derivatisation allows the selective enhancement of the b-ion series in a MS/MS spectrum. The application of the ion labelling technique is demonstrated by various examples. The derivatisation procedure results in a characteristic tag. The modification of the spectra could be used for automated interpretation of mass spectra and protein quantitation by stable isotope labelling. Parts of this chapter will be published in the following articles:


4.1 Chemical Identification

4.1.1 N-terminal

Several methods exist for the identification of the N-terminal residues of a protein. In 1952, Frederick Sanger (Sanger, 1952) deciphered the amino acid sequence of insulin. He reacted the compound 1-Fluoro-2,4-dinitrobenzol with the N-terminal group of a polypeptide in alkaline solution. The polypeptide was subsequently hydrolysed and the N-terminal amino acid was labelled with the dinitro benzene group, which imparts a bright yellow colour. The amino acids were separated and the N-terminal residue was identified by comparison with the migration of known dinitrophenyl derivatives.

If one thinks today of sequencing a protein, one (still) will choose the method developed and automated by Pehr Edman in 1949 (Edman, 1949; Edman and Begg, 1967). In this method, the N-terminus of a polypeptide is labelled and released from the polypeptide. The cyclic reaction can be repeated over a hundred times, though in general a standard length is around 30 amino acids. In alkaline solution, phenylisothiocyanate (PITC) is reacted with the terminal amino group to yield a phenylthiocarbamyl derivative of the peptide. This derivative is then treated with a strong anhydrous acid (100% TFA), which results in cleavage of the peptide bond between residues 1 and 2. The derivative of the N-terminal residue is then transferred to a separate vessel where it is treated with 25% TFA in water. It rearranges to yield a phenylthiohydantoin (PTH) derivative of the amino acid. This PTH-amino acid derivative is subsequently identified by a comparison of the retention time with that of the standard PTH-amino acid derivatives. Two important things have been accomplished. First, the N-terminal residue has been marked with an identifiable label, as in the Sanger method. However, in the Edman method, the rest of the polypeptide has not been destroyed; it has simply been shortened by one residue, so that the whole set of reactions can now be repeated and the second residue determined. Even after 50 years, the general chemistry is still the same and great improvements in the instrumentation and detection systems have made this method a useful tool for the identification of peptides, down to the 400 femtomole level. The development of gas-phase sequencing from non-covalent supports made it possible to sequence PVDF-immobilised proteins isolated directly from SDS-PAGE (Aebersold, 1990) and 2D-PAGE gels.
4.1.2 Ladder sequencing

Ladder sequencing, in which a sequence is read by the mass difference between sequential degradation products, was developed first as an enzymatic technique (Aimoto et al., 1982) and subsequently as a modified Edman-type chemical degradation (Chait et al., 1993). The use of exopeptidase is limited by the extreme variability of the activity of the protease towards the substrate and is only useful for individual isolated peptides (Korostensky et al., 1998).

The main problem with the traditional Edman approach is, that the sequencing is slow, setting the limit at ~ 50 residues per day. To overcome this, Brian Chait modified the old Edman method by combining multiple steps of wet degradation chemistry with a final, single step mass spectrometric read-out by MALDI-TOF-MS.

Figure 4-1  Protein ladder sequencing (Chait et al., 1993) exemplified by the generation of a set of sequence-determining fragments from an intact peptide chain with controlled ladder generating chemistry. A stepwise degradation is carried out with a small amount of terminating agent present in the coupling step. In this case, 5% phenylisocyanate (PIC) was added to the phenylisothiocyanate (PITC). The phenylcarbamyl (PC) peptides formed are stable to the acid cleavage by TFA used to cyclise and cleave the terminal amino acid from the phenylthiocarbamyl (PTC) peptide. Successive cycles of ladder generating chemistry are performed without intermediate isolation of released amino acid derivatives. Finally, the mixture of PC peptides is read out in one-step MALDI-TOF MS.
First, a sequence-defining concatenated set of peptide fragments, each differing from the next by a single residue, is chemically generated in a controlled fashion. The standard Edman reagent phenylisothiocyanate (PITC) is used to carry out stepwise degradation. A small amount of terminating reagent is added to the coupling mixture (phenylisocyanate, PIC) to block a small amount of amino termini at each cycle. The phenylcarbamyl peptides formed are stable to the TFA cleaving step. Successive cycles of ladder-generating chemistry are performed without intermediate isolation or analysis of released amino acid derivatives. The resulting mixture is read out in a single MALDI-TOF-MS experiment. The mass differences between consecutive peaks each correspond to an amino acid residue and their order of occurrence in the data set defines the sequence of amino acids (see Figure 4-1). The main disadvantages of this method are the loss of peptide during washing steps, which limits the sensitivity, as well as the terminating reagent, which removes the alpha N-terminus as a charge carrier thereby diminishing detectability by MS. It is not applicable to intact proteins >30kDa and only a single peptide should be present.

A different method for generating multiple small sequences from the N-terminal of peptides in unseparated protein digests by stepwise thioacetylation and acid cleavage was developed in our laboratory (see Figure 4-2). In this approach, the sequence ladders are generated without the use of chain terminators or sample aliquoting. The degradation reagents are water-soluble so that the chemistry can be carried out on peptides immobilised on C18 reverse phase supports without any peptide loss due to washing with organic solvents as occurs in Edman-type sequencing. The entire chemical tagging method can be automated and a prototype device for parallel analysis of multiple samples was constructed (Hoving et al., 1999).

In this study, thioacetylthioglycollic acid (TATG) is used as a coupling reagent. The reagent is coupled in the presence of N-ethylmethylpiperidine and the chemical tagging procedure is carried out as described in Figure 4-2.
Figure 4-2  A General outline of the Chemical tagging procedure. Coupling of thioacetylthioglycolic acid in alkaline conditions to the N-terminal part of the peptide or protein followed by cleavage of the first amino acid residue by strong acid (TFA) and release of the thiazoline. B Principle of ladder sequencing. A schematically mass spectrum of a tryptic protein digest before and after three degradation cycles.
4.2 **Tandem Mass Spectrometry**

The advent of tandem mass spectrometry made it possible to identify peptides in complex mixtures without the need to separate the individual components (Hunt et al., 1981). A mass window narrow enough to exclude all but one of the mixture's components can be chosen in the first of the two mass spectrometers of a tandem instrument. In this step, the precursor ion is selected and in a following step dissociated. The dissociation is usually induced by collision with an inert gas (Argon or Xenon; collision induced dissociation, CID). CID produces a product ion spectrum in which all fragment ions are derived from the selected precursor ion (see Figure 4-3). Using this technique, it is possible to identify and sequence peptides from unfracti onated enzyme digests of proteins or from digests partially fractionated by HPLC.

![Figure 4-3 Schematic depiction of the process of acquiring tandem CID mass spectra of peptides in a triple quadrupole mass spectrometer](image)

**4.2.1 Tandem-in-space and tandem-in-time mass spectrometry**

A triple quadrupole MS instrument as depicted in Figure 4-3 is essentially a set of three quadrupoles in series. This instrument is tandem-in-space, in the sense that ion reactions studied are isolated from possible interfering reactions by spatial dispersion of the appropriate ions. Thus for each reaction step, such instruments require a separate and well-defined spatial region. This is in contrast to tandem-in-time instruments, in which sequential steps are studied in the same physical space, but separated from one another via an appropriate timing sequence of excitation and detection. An ion trap is a typical example of such a device.
4.2.2 Low energy peptide fragmentation

One of the first groups to sequence underivatised peptides by tandem mass spectrometry was Don Hunt (Hunt et al., 1986). He managed to exploit low-resolution parent ion selection and daughter ion detection to maximise sensitivity and chemical modification to aid the determination of ion series for interpretation. The protonated parent ion undergoes multiple low energy collisions with the collision gas. The kinetic energy is then converted into vibration energy and fragmentation at one of the amide bonds occurs. The generally accepted nomenclature for ions is that proposed by Roepstorff & Fohlman (Roepstorff and Fohlman, 1984) and modified by Biemann (Biemann, 1988) (see Figure 4-4).

The amount of energy transferred in a collision is very small, ca. 1% tailing up to 15%. The quadrupole and ion trap instruments produce “low energy” collisions of around 10-30eV. The energy of a peptide bond is approximately 1eV (1eV is 1.6 x 10^-19 J). Although the energy transferred is only of the order of 0.1eV, peptide bond cleavage occurs since the gas pressure in the collision cell is set such that the ions undergo multiple collisions and can accumulate the necessary energy. The effectiveness of fragmentation is dependent on the number of collisions (gas pressure), the energy deposition (the greater the cross-sectional area of the gas molecule the larger the transfer, i.e. Xe > Ar > He) and the charge on the ion. The collision energy is set in a triple quadrupole and quadrupole-TOF according to the charge and mass of an ion. About 50eV for a singly charged ion of mass 2000, or the 2⁺ ion of mass 1000 etc. with a constant gas.
pressure of 2.5mTorr Argon is needed to induce fragmentation. The ion trap operates under a constant pressure (1-2mTorr) of helium and collisional activation is brought about by causing the ion to vibrate at its resonant frequency causing it to undergo multiple low energy collisions.

The fragmentation mechanism for CID is not well understood, but empirically, the following principles can be given (Hunt et al., 1986). In the liquid or solid phases used for ESI and MALDI, the peptides exist as cations with positive charges in the form of protons localised on the amino-terminus and basic side chains such as arginine, lysine and histidine. The acidic residues, glutamic and aspartic acid are protonated and hence neutral. Solvation delocalises the charge and stabilises the charged species. In the gas phase in the absence of solvent molecules, charge transfer occurs by direct contact between carriers. Since the gas phase proton affinity of the peptide backbone amide linkage is equivalent to that of a primary amine, chain folding serves to allow free movement of protons from the amino terminus and lysine side chains. This results in an even distribution along the backbone.

In a triple quadrupole instrument the peptide ions selected by the first MS reach the collision cell. There they undergo multiple (1-10) collisions with the argon atoms and the kinetic energy is converted into vibrational energy. The energy is then sufficient to cleave one of the amide backbone bonds. The subsequent fragmentation pattern depends on the collisional energy, the pressure and type of the collision gas, the number of charges being carried by the peptide, as well as the amino acid sequence of the peptide. If the collision energy is relatively high, then fragmentation involving a single bond cleavage is favoured and b-ions are preferentially formed. However if the energy is relatively small, a pathway with lower activation energy is favoured involving a simultaneous bond breakage and formation, and y-ions are preferentially formed. If the peptide is singly charged then one of the daughter ions formed is charged and the other is neutral. The neutral molecule does not respond to Rf and is lost.

In an ion trap instrument, the parent ion is selected by using U and Rf fields to eject all ions except the selected. The collisionally induced dissociation is achieved by using resonance excitation to increase the kinetic energy of the trapped ion (though keeping the resonance amplitude low enough to prevent ejection). The trapped ion undergoes multiple collisions with the helium bath gas, which is continually present in the trap. The mass spectrum of the daughter ions is recorded using a mass-selective instability scan in which the amplitude of the Rf voltage applied to the trapping ring is scanned. This causes ions of increasing m/z values to adopt unstable trajectories and to be ejected from the trap through the end cap, where they are detected by an external electron multiplier. At this stage, a daughter ion of interest can be kept in the
trap and again forced to dissociation using resonance excitation. Thus, in a tandem-in-time instru-
ment as an ion trap instrument, multiple dissociation experiments can be carried out (MS\textsuperscript{n} experi-
ments).

The distribution of charge plays an essential role in determining the relative amounts of b and y ions that are detected. This is also dependent on the number and position of Pro, His, Trp, Arg and Lys residues, which have a high basicity in the gas phase. In the low energy regime of a triple quadrupole described here, the neutral and charged products do not separate immediately after bond cleavage but remain in close contact by dipole-charge attractions until further collisions cause separation. During this time, proton transfer to the residue with the highest basicity can occur, hence explaining the relative dominance of y ions in tryptic peptides that usually have Arg or Lys as the terminal residue. This is not the case for the products in an ion trap since only the parent ions are excited to resonance and hence to energy transfer collisions, the products are not. Hence there are proportionately more b ions found in a spectrum of a peptide obtained using an ion trap MS than with the same peptide in a triple quadrupole MS.

Often ion types are found which have eliminated small molecules as carbon monoxide (-28 U; b-ion to an a-ion transformation), ammonium (-17 U) and water (-18 U). The loss of a water molecule occurs at serine or threonine residues and the corresponding ion series is labelled with \(\circ\), indicating the loss). One of the major limitations of low energy fragmentation is that the isobaric amino acids Isoleucine and Leucine cannot be distinguished.

The interpretation of such tandem mass spectra is the slow step in the determination of the amino acid sequence of peptides by mass spectrometry, especially when compared with the Edman sequencing method. Progress has been made with automated interpretation programs (see chapter 5), but they are still far from being general usable, so a manual re-interpretation of computer-interpreted spectra is always necessary. The interested reader in interpretation of peptide MS/MS spectra is referred to the detailed review from Papayannopoulos (Papayannopoulos, 1995).

\textit{De novo} sequencing by mass spectrometry is a difficult task. In a normal MS/MS fragmentation spectrum, the main difficulty is the identification of the produced internal ions. The assignment of all the N-terminal ions (a, b, c and b\textsuperscript{2+}) and C-terminal ions (x, y, z and y\textsuperscript{2+}) to the different peaks is far from trivial. The spectrum gets even more further complicated, when internal ions are produced, which mostly but not always happens next to proline and histidine residues. Nevertheless, sequencing by MS/MS is a powerful tool, especially when combined
with database search programs as SEQUEST and the complete genome sequence of the organism (see chapter 5). If the genome sequence of the organism is available, from which the protein of interest is obtained, manual interpretation of mass spectra is the last and only resort.

In a usual proteomic experiment, the organism of interest is grown under two different conditions. Either an external parameter such as heat, nutrition, growth factors, morphogens or oxygen supply is modulated or an internal parameter is changed by gene knock-out or by the introduction of plasmids carrying regulators or additional genes. The proteome is visualised by 2D PAGE as described in chapter 2. The standard processing procedure is depicted in Figure 4-5:

![Diagram](Figure 4-5 General processing scheme. The spot of interest is picked from several 2D gels and concentrated in a funnel-shaped gel device. The concentrated spot is digested by the protease of choice and the produced peptides are extracted from the gel. The digest is desalted over reverse-phase column and subsequently analysed by mass spectrometry.)

4.2.2.1 Concentration gels

A modified version of the funnel concentration gel described by Lombard-Platet and Jalinot (Lombard-Platet and Jalinot, 1993) was used (see Figure 4-5). The gel consists of a long funnel-shaped stacking gel (4% acrylamide) and a short resolving gel (12% acrylamide). Coomassie® Blue G250 stained spots were cut out from 2D-PAGE gels, washed in distilled water (3 x 15 min) and solubilised in the standard SDS-containing Laemmli solubilisation buffer. Electrophoresis was carried out at constant current of 10mA using bromophenol blue to track the progress of the run. In the OGP-PAGE version, SDS was substituted by 0.1% β-octylglucopyranoside in the running and stacking gel buffers and no detergent was used in the electrode buffer. The gel was finally stained with Coomassie® Blue G250 as usual.
4.2.2.2 Proteolytic in-gel cleavage of peptides

In order to generate peptides small enough to carry out MS/MS studies, the concentrated protein spot is in-gel-digested (Rosenfeld et al., 1992; Shevchenko et al., 1996). The gel piece is extensively washed with water to remove all the excess acetic acid from the previous staining procedure. The gel is cut in little pieces to increase the surface area of gel material, which is subsequently dehydrated in pure acetonitrile for 15 minutes. The particles are brought to complete dryness in a vacuum centrifuge. The gel is reswollen in 10mM dithiothreitol / 0.1M NH₄HCO₃ and incubated for 30 minutes at 56°C to reduce the protein. After the excess liquid is removed, the particles are shrunk in acetonitrile. The solvent is replaced with 55mM iodoacetamide / 0.1M NH₄HCO₃ and the sample incubated for 20 minutes at room temperature. After the removal of iodoacetamide solution, the gel is shrunk in acetonitrile and in a vacuum centrifuge. If the gel particles still contain Coomassie® Blue staining, the particles are rehydrated in 0.1M NH₄HCO₃ and after 15 minutes reshrunk in acetonitrile. This step is repeated several times until the stain is completely removed. This is especially important, if the sample is to be analysed in an electrospray mass spectrometer and no previous HPLC purification is done. Now the sample is ready for in-gel digestion. The gel is rehydrated in the digestion buffer containing 50mM NH₄HCO₃ and 12.5ng/μl trypsin at 4°C. After 15-20 minutes, the sample is checked and more buffer is added, if all liquid is absorbed by the gel pieces. After an additional 25 minutes the remaining supernatant is removed and replaced by the same buffer but without trypsin to cover all the gel pieces. The digestion is carried out over night at 37°C on an Eppendorf shaker. To analyse the digest by ESI MS/MS, the tryptic peptides have to be extracted from the gel particles. The supernatant is collected and subsequently replaced by 25mM NH₄HCO₃. After incubation for 30 minutes at 37°C, the supernatant is collected and replaced by acetonitrile for 15 minutes. After collecting and replacement by 5% formic acid, the particles are vortexed for 15 minutes. In a last step, the particles are shrunk in acetonitrile and all the collected supernatants reduced to a volume of approximately 10μl in a vacuum centrifuge.

4.2.3 Derivatisation – Simplifying Interpretation

To sequence a protein, which is not in the database, is not straightforward and fast. It’s possible to obtain a protein sequence and even to see post-translational modifications, but the interpretation of such spectra takes even an experienced person a reasonable amount of time. To ad-
dress this problem and make the interpretation of mass spectra faster and more reliable, two new approaches will be described in this part.

4.2.3.1 Ion labelling I: The $^{16}\text{O}/^{18}\text{O}$ method

One method has been used in mass spectrometry for many years and was adapted by Rose and colleagues (Rose et al., 1983) for peptide sequencing. Rose carried out a proteolytic digest in buffers prepared with a 50:50 mixture of $\text{H}_2^{16}\text{O}/\text{H}_2^{18}\text{O}$. Because of the proteolytic digest, all C-terminal fragments of all generated peptides except the C-terminal peptide are characteristically marked by a doublet peak resulting from the difference of two mass units between $^{16}\text{O}$ and $^{18}\text{O}$ (see Figure 4-6). This doublet marking is very useful, if a labelled peptide is subsequently fragmented in a CID experiment. Consequently, all y-ions appear as doublets and make the interpretation of mass spectra much easier.

The protein of interest was either in-gel digested as described in 4.2.2.2 or digested in solution as described in 4.2.3.2. To label the produced peptide fragments the enzymatic cleavage solution contained 50% $\text{H}_2^{18}\text{O}$. Because of the high costs of $\text{H}_2^{18}\text{O}$ ($1000$ per ml), the volumes were kept as small as possible. For a single spot excised from a 2D gel, a minimum volume of 25$\mu\text{l}$ is needed. The composition of the digestion mixture was as follows (3$\mu\text{l}$ 0.1$\mu\text{g/\mu l}$ Trypsin, 12$\mu\text{l}$ $\text{H}_2^{16}\text{O}$, 10$\mu\text{l}$ $\text{NH}_4\text{HCO}_3$). The digest was analysed by MALDI TOF-MS and subsequently ESI MS/MS analysis was performed (see 4.2.3.2). The digested peptides were dried in a speed-vac to remove the volatile buffer. The sample was taken up in 50% methanol and 1% acetic acid and dried again. The digest was resuspended in 0.1% TFA and desalted over the same buffer prior to analysis by ESI MS/MS.
4.2.3.2 Ion labelling II: Nicotinic Acid derivatisation

We have developed a complementary approach to isotopically label the b-ion series resulting from a CID experiment. The fragments generated by a protease are labelled by a 50:50 mixture 1-(H₄-nicotinoyloxy) succinimide and 1-(D₄-nicotinoyloxy) succinimide as shown in Figure 4-7. The synthesis of the 1-(H₄/D₄-nicotinoyloxy) succinimide (H₄/D₄-NA-NHS ester) was performed as outlined in Figure 4-8. 0.25g nicotinic acid was dissolved in 5ml tetrahydrofuran (THF) and mixed with 0.4g dicyclohexylcarbodiimide under continuous stirring in a reaction flask for 2 hours at room temperature. In the next synthesis step, 0.25g N-hydroxysuccinimidyl ester were added to the solution and stirred over night at room temperature. The yellowish precipitate was filtered and the solution crystallised in ethylacetate. The purity of the product was checked by ¹H-NMR.

¹H-NMR (CDCl₃, 300 MHz):

\[ \delta = 2.863 \text{ (s, 2 CH₂), 7.40-7.50 (m, H₅ (ArH)), 8.30-8.40 (dt, H₄ (ArH), J = 1.86, 2.18, 8.1 Hz), 8.80-8.90 (dd, H₆ (ArH), J = 1.87, 4.93 Hz), 9.27 (d, H₂ (ArH), J = 2.18 Hz) ppm.} \]

The succinylation was performed according to Hui, J.O. et al. (Hui et al., 1990) as follows. Myoglobin (1mg) or a protein of interest is dissolved in 1ml 100mM N-ethylmorpholine acetate (pH 8.5) containing 6M guanidinium hydrochloride, and the solution was treated with a
Chapter 4 Identifying Proteins

Figure 4-7  General outline of the $H_2D_2$-Nicotinic Acid-N-hydroxysuccinimid ester method. A protein or peptide of interest is tryptically digested and subsequently modified in a 50:50 mixture of $H_2D_2$ Nicotinic Acid-N-hydroxysuccinimid ester. Therefore, 50% of the resulting 'new' N-termini are labelled by $H_4$ Nicotinic Acid and 50% by $D_4$ Nicotinic Acid, thus creating a visible tag for the b-ions in MS/MS spectra.

Figure 4-8  Scheme of the synthesis of 1-(nicotinoyloxy) succinimide ($H_2D_2$-NA-NHS ester). Nicotinic Acid reacts with dicyclohexylcarbodiimide to yield 1-(nicotinoyloxy)-N, N'-dicyclohexylurea. The urea derivative reacts with N-hydroxysuccinimide, which results in 1-(nicotinoyloxy) succinimide and N, N'-dicyclohexylurea.

200-fold molar excess of solid succinic anhydride. The anhydride was added over a period of 1h, and the pH was maintained at 8.5 by the dropwise addition of 1N NaOH. After the last addition of reagent, the reaction mixture was allowed to stir at room temperature for 1h. The clear solution of succinylated protein was purified by reverse-phase HPLC on a capillary column (Macherey-Nagel, ET 250/2 Nucleosil 300-5 C18, 250 x 2 mm; ABI 140B solvent delivery system, ABI 1000S-diode array detector).
The purified fractions from the reverse phase HPLC run were dried in the speed-vac and redissolved in 50μl 100mM NH₄HCO₃. *Staphylococcus* protease V8 (Promega, sequencing grade) was added at 2% w/w with respect to the succinylated protein. Digestion was performed over night at 37°C. The digest was dried in the speed-vac to remove the volatile buffer, washed with 20ml water, dried again and finally the peptides were redissolved in 50μl 100mM NH₄HCO₃. A second digest was performed by adding porcine trypsin (Promega, sequencing grade) at 2% (w/w) with respect to the succinylated peptides. Digestion was carried out 6h at 37°C. The digest was dried in the speed-vac to remove the volatile buffer, washed with 20μl water, dried again and finally the peptides were redissolved in 50μl 500mM N-ethylmorpholine acetate (pH 8.5).

The peptide digest mixture was derivatised by adding a 50:50 mixture 1-((H₄-nicotinoyloxy) succinimide and 1-((D₄-nicotinoyloxy) succinimide at 100-fold molar excess with respect to the starting protein concentration. The reaction was performed on an Eppendorf shaker (Thermomixer 5436) at 37°C for 45 min and the reaction checked on MALDI-TOF-MS.

0.5μl of crude digestion mix was co-crystallised with the same amount of matrix solution (10mg α-cyano-4-hydroxy-cinnamic acid in 50% acetonitrile and 1.25% TFA in water). The sample-matrix mix was allowed to air dry and subsequently washed three times by adding ice-cold 1% TFA to the mixture on the MALDI target. Mass spectra were recorded using a Voyager Elite MALDI-TOF mass spectrometer (Perseptive Biosystems, Framingham, MA, USA). Samples were analysed in delayed extraction reflector mode using an accelerating voltage of 20kV, a pulse delay time of 150ns, a grid voltage of 60% and a guide wire voltage of 0.05%. Spectra were accumulated for 32 or 64 laser shots.

Before analysing the succinylated, V8 and tryptic digested and nicotinylated peptide mix by ESI-MS/MS sequencing, the complete mixture was purified by reverse-phase HPLC on a capillary column (Macherey-Nagel, ET 250/2 Nucleosil 300-5 C₁₈, 250 x 2 mm; ABI 140B solvent delivery system, ABI 1000S diode array detector). The purified fractions from the reverse phase HPLC run were dried in the speed-vac and redissolved in 50% methanol and 1% acetic acid before introduction into the mass spectrometer at a flow rate of 0.2μl/min with a syringe pump. MS/MS sequencing was performed on a Finnigan MAT LC-Q ion trap mass spectrometer (San Jose, CA, USA). The peaks of interest were selected with a mass window of ± 3 a.m.u. Fragmentation was established using a relative collision energy of 35-60 units for MH⁺-ions and 20-35 units for MH²⁺ ions.
4.3 Application of the Techniques

In the following part of this thesis, only sequencing by mass spectrometry is covered, though Edman sequencing was routinely used to determine N-termini of proteins in our laboratory.

4.3.1 MS/MS Sequencing

Many nitrogen fixation-associated genes in the soybean symbiont *Bradyrhizobium japonicum* are associated with characteristic -24/-12 promoters, whose recognition by RNA polymerase requires a σ^54 factor (Merrick, 1993). These σ^54 promoters are dependent on transcriptional activator proteins, which are modulated in response to specific environmental signals (Shingler, 1996). The transcriptional activator for nitrogen fixation genes is NifA (Fischer, 1994), whose activity is inhibited by aerobiosis. NifA is encoded in the *fixR-nifA* operon, which is expressed at a low level under aerobic conditions and induced approximately fivefold under low-oxygen tension. This induction depends on a -24/-12-type promoter (*fixRp1*) that is recognised by the σ^54 RNA polymerase and activated by NifA. Low-level aerobic expression and part of the anaerobic expression originates from a second promoter (*fixRp2*) that overlaps with *fixRp1* and depends on an upstream DNA region (upstream activator sequence, UAS) located around position -68 (Barrios et al., 1995). RegR was identified to act as an activator by binding to the UAS (Bauer et al., 1998). Based on the predicted amino acid sequence, RegR belongs to the family of response regulators of two-component regulatory systems. Two component regulatory systems are one of three different prokaryotic systems that signal by protein phosphorylation (Cozzone, 1998). The system uses a reversible modification to regulate cellular activities. Initially a sensor kinase is autophosphorylated at a histidine residue in response to the activation signal. The phosphor group is transferred onto an Asp residue of the response regulator and then binds the target inducing a change in gene expression. Bauer et al. identified upstream of the *regR* gene an additional gene (*regS*) which encodes a putative sensor kinase. A *regR* mutant was constructed in which neither a specific UAS-binding activity nor *fixRp2*-dependent transcript formation and *fixR*-lacZ expression was detected in aerobically grown cells. The newly identified RegR protein may be regarded as a master regulator in the NifA-dependent network controlling *nif* and *fix* gene expression in *B. japonicum*.

A proteome approach was chosen to identify additional targets of the postulated master regulator. *B. japonicum* wt and *regR* mutant cells were grown aerobically and a batch of 20 2D-
PAGE gels was run. Comparative analyses of the obtained gels revealed an interesting new target present in the wt gel and not present in the regR mutant. This spot was picked from the wt gels and concentrated in the funnel-gel device as described in 4.2.2.1. The concentrated band was cut in two halves. One part was electroblotted onto PVDF membrane in a semidry apparatus. The proteins blotted onto PVDF membrane were N-terminal sequenced on a Hewlett Packard G1000A. Edman sequencing revealed no amino acid sequence, probably due to N-terminal blocking of the first amino acid. An artificial chemical blocking of the N-terminal amino acid during the experimental procedure can be excluded, because the internal control standard BSA, processed in parallel with the spot of interest, gave a clear signal in the Edman sequencing reaction.

The other half was in-gel digested by trypsin, extracted and purified on reverse-phase as described in 4.2.2.2. It was possible to obtain three internal sequences by MS/MS analysis in the ion trap from the spot of interest:

\[
\]

![Figure 4-9 2D PAGE gels of B. japonicum wt A and regR mutant B. Whole cell extract was resolved on a pl 3-10 in the first dimension and subsequently separated according to the molecular weight in a 12% SDS-PAGE. The spot circled in A is not present in the regR- mutant. This spot was excised, concentrated and digested as described in the text.](image)

The MS/MS spectrum of the peptide S-V-S-R-K is shown in Figure 4-10, as an example of peptide sequencing by mass spectrometry. The peptide is derived from a tryptic digested protein spot, picked from a 2D-SDS PAGE (pl 3-10) of the soluble proteins from the regR mutant of B. japonicum. The sequence will be used in further studies to screen B. japonicum for the coding gene of the spot picked from the 2D gel. Because of the non-appearance in the mutant gel, the appearing spot in the wt gel must be RegR dependent and should be under the control of the RegR regulon.
Figure 4.10 An example of MS/MS spectra interpretation. **A** a zoom scan of a single peptide derived from a tryptic digest of a protein spot picked from a 2D-SDS PAGE from the regR mutant of *Bradyrhizobium japonicum*, the nitrogen fixing root-nodule symbiont of Soybean is shown. **B** the same peptide is forced to fragment in an electrospray ion trap instrument, producing a characteristic pattern from which the amino acid sequence SVSRK can be derived.
4.3.2 Derivatisation

As one can see from Figure 4-10 the interpretation of an MS/MS spectrum is very time consuming and needs considerable expertise. To simplify the interpretation of a spectrum it is very helpful to selectively label an ion series. A label on the N-terminus of a peptide, for example, allows easy identification of the N-terminal or b-ions by their characteristic isotope pattern. The higher data quality obtained will enable automated interpretation of the spectra by a simple software algorithm. This technique will provide a step to routinely sequence by mass spectrometry.

4.3.2.1 Derivatisation chemistry I: The Nicotinylation

Specifically, the peptides obtained from a protein digest are derivatised by reacting with a 50:50 mixture of l-(H4-nicotinoyloxy)succinimide and l-(D4-nicotinoyloxy)succinimide. Subsequent fragmentation by MS/MS gives a b-ion with a characteristic 4 mass unit shift, which simplifies the identification of the b-ion series.

In an initial experiment, a tryptic digest of myoglobin was derivatised at a range of different pHs. The modification of the tryptic myoglobin digest in 500mM NEM (pH 5.5) resulted in a selective labelling of the N-termini of all peptides obtained. When the same experiment is performed in 500mM NEM (pH 8.5), ε-amino groups at lysine residues are modified as well. This can be explained that at pH 5.5 α-NH₂ groups non protonated and reactive, while ε-NH₂ groups are protonated and non reactive. On the other hand, at pH 8.5 α-NH₂ as well as ε-NH₂ groups are non protonated and reactive, resulting in the modification of both types of amino groups in the peptide. Because of the restriction of the serine protease Trypsin on lysine or arginine residues as cutting site, there is always a C-terminal residue with high gas phase basicity. When a peptide is derivatised at all it’s amino groups, there is a label attached N- and C-terminal, marking both ends of the peptide chain. This results in a labelling of the b- and y-ion series and does not simplify the interpretation of spectra.

As one can see from Figure 4-11, the pH of the nicotinylation reaction has a great influence of the amino group modified. At pH 5.5 a fast modification of the N-terminal amino group was found. When longer reaction times were applied (> 30min) a modification of the ε-amino groups from internal lysine residues are nicotinylated was found as well. At basic conditions (pH 8.5) a fast and complete nicotinylation of all amino groups was found (~20min).
Chapter 4

IDENTIFYING PROTEINS

A

![Graph and sequence analysis]

1271 (A): L-F-T-G-H-P-E-T-L-E-K
1663 (D): L-F-T-G-H-P-E-T-L-E-K-P-D-K

B

![Graph and sequence analysis]

1271 (A)
1379 (B)
1607 (C)
1663 (D)
1817 (E)
1854 (F)
1886 (G)

1376 (A')
1484 (B')
1712 (C')
1768 (D')
1922 (E')
1959 (F')
1991 (G')

pH 5.5

+ Nicotinic Acid - NHS
Identifying Proteins

Chapter 4

Figure 4-11 MALDI-TOF MS of myoglobin, trypically digested: the difference in the reactivity of 1-(H4-nicotinoxyloxy)succinimide at pH 5.5 and pH 8.5 is shown. In A unmodified myoglobin is digested by trypsin and the trypic fragments are seen in the spectrum. In B and C the trypic digest is modified at pH 5.5 and pH 8.5, respectively. Under acidic conditions, only the a-amino groups are labelled by nicotinic acid, while under basic conditions a-amino groups as well as e-amino groups from lysine residues are nicotinylated.

Mass (m/z)

1271 (A) - 1376 (A') - 1481 (A'')
1379 (B) - 1484 (B') - 1589 (B'')
1607 (C) - 1712 (C')
1663 (D) - 1768 (D') - 1873 (D'')
1817 (E) + Nicotinic Acid - NHS
1854 (F) - 1959 (F') - 2064 (F'') - 2169 (F''')
1886 (G) - 1991 (G') - 2096 (G'')

In Figure 4-12 the MS/MS spectrum of a myoglobin trypic digest peptide is shown. The peptide was derivatised with a 50:50 mixture of 1-(H4-nicotinoxyloxy)succinimide and 1-(D4-nicotinoxyloxy)succinimide at different pH-values. In A the b-ion series from b4 to the parent ion is clearly visible. With the help of the multiple MS-spectra MS3 and MS4 displayed in B and C respectively, the complete sequence of the peptide can be revealed. In D the mass difference between b140 (parent ion –18u, due to the loss of a water molecule) and b13 of 233 a.m.u. can be explained by the loss of 1-(H4/D4-nicotinoxyloxy)lysine (128.09 + 105/109). The b-ion series b4 to b13 appear at the same m/z value, due to the cleavage of the C-terminal modified lysine. An internal lysine would have shifted the whole b-ion series up to the modified lysine.

In this particular peptide H-G-T-V-V-L-T-A-L-G-I-L-K, the signal of the b-ion series is enhanced due to the N-terminal histidine residue, being a charge acceptor and the N-terminal modification by the pyridine derivative, explaining the reduced intensity of the y-ion series.

4-21
CHAPTER 4

IDENTIFYING PROTEINS

A: MS/MS 1484/1488
myoglobin; tryptic digest; H4/D4 nicotinylated pH 5.5

B: MS/MS 1484/1488
myoglobin; tryptic digest; H4/D4 nicotinylated pH 5.5

C: MS/MS 1484/1488
myoglobin; tryptic digest; H4/D4 nicotinylated pH 5.5
The phenomenon, that the distribution of product ions between the two species varies is known (James, 1997). The distribution is very much dependent on the number and distribution of the residues with high gas phase basicity (lysine, arginine, proline, tryptophane and histidme). The neutral and charged fragments produced by low energy fragmentation do not immediately separate following bond cleavage. They remain associated and undergo further collisions, so proton transfer to the fragment of higher basicity may explain the preferential distribution of fragments from a certain ion series. In MS/MS spectra of a tryptic digest containing peptides, which have the basic amino acids arginine or lysine as C-terminal amino acid, mostly the \( \gamma \)-ions are predominant.

With an N-terminal modification and histidine residue, it is possible to enhance the \( b \)-ion series as shown in Figure 4-12. The modification of methionine-Enkephalin (\( Y-G-G-F-M-B \), \( B = \) nicLys) at the N-terminal amino group and the C-terminal lysine results in the labelling of both ion series. This results in an enhancement of the \( b \)- and \( \gamma \)-ion series as shown in Figure 4-13A.
Both series appear with the same relative abundance and the discrimination between both series is not as obvious as in Figure 4-12. No other amino acid with high gas phase basicity as histidine, proline and tryptophan is present to influence the abundance of one ion-series. The interpretation of the spectrum is as complicated as without labelling. The unmodified spectrum of Met-Enkephalin (Y-G-G-F-M-K) can be seen in Figure 4-13 B.

A: MS/MS 912/916/920
Met-Enkephalin, D_4/nicotinylated

B: MS/MS 702
Met-Enkephalin-Lys, native

Figure 4-13  MS/MS spectra of Met-Enkephalin derivatised with 50:50 mixture 1-(H_4-nicotinoxyloxy)succinimide and 1-(D_4-nicotinoxyloxy)succinimide. A shows Met-Enkephalin derivatised at pH 8.5. The b-ion series as well as the y-ion series is enhanced due to the label attached at the N-terminal amino group and the C-terminal ε-lysine amino group. B Spectrum of unmodified, native Met-Enkephalin.
The effect of residues with high gas phase basicity is obvious in Figure 4-14. The peptide A-D-I-A-G-H-G-Q-E was obtained by digesting succinylated myoglobin with V8, then trypsin and subsequent labelling by 50:50 mixture 1-(H₄-nicotinoyloxy) succinimide and 1-(D₄-nicotinoyloxy)succinimide. As one can clearly see, the b-ion series is marked with its characteristic double peak of a mass difference of four a.m.u. The y-ion series is prominent due to the presence of histidine, but appears as normal peaks. Thus, the difference of the b- and y-ion series is clearly visible and the interpretation of the spectrum simple.

Figure 4-14 MS/MS spectra of myo peptide A-D-I-A-G-H-G-Q-E derivatised with a 50:50 mixture of 1-(H₄-nicotinoyloxy)succinimide and 1-(D₄-nicotinoyloxy)succinimide. A b-ion marked by the characteristic double peak, y-ion series is enhanced due to the presence oh the histidine residue. B MS² of the b⁴ ion group from myo peptide A-D-I-A-G-H-G-Q-E
4.3.2.2 Derivatisation chemistry II: The Succinylation

As was shown in Figure 4-13, the derivatisation by nicotinic acid is pH dependent and even under acidic conditions not completely restricted to α-amino groups. To avoid a reaction with internal lysine residues, resulting in the labelling of both ion series, the protein was modified by succinylation (Figure 4-15). The labelling of lysine residues is helpful in an additional way, because it allows one to discriminate between lysine (128.09) and glutamine residues (128.05). Succinic anhydride selectively modifies amino groups in peptides, thus modifying N-termini and internal lysines. Non-digested myoglobin was modified by succinic anhydride and purified by reverse-phase HPLC. Therefore, the succinylated lysine residues are not recognised by the protease trypsin and only arginine serves as a cutting site. The derivatised protein was subsequently digested by endoproteinase Glu-C (V8), which cuts C-terminal to aspartic and glutamic amino residues.

Figure 4-15 Succinylation of myoglobin. A shows ESI-ion trap MS with the multiple charged positive ions. B the deconvoluted spectrum of myoglobin using the BIOMASS deconvolution algorithm. C HPLC chromatogram at 280nm wavelength, 20μg myoglobin injected. D HPLC chromatogram at 280nm wavelength, 680μg myoglobin injected. In both HPLC runs a Rp-8 cartridge (Brownlee Labs Inc, Santa Clara CA, USA), 30 x 2.1μm, 5μm was used with the following gradient: 0%B for 10 min, 0-100%B in 15 min, 100%B in 2 min.
Figure 4-16  MALDI-TOF MS of myoglobin, trypic digested: In A succinylated myoglobin is digested by GluC-protease (V8). In B the digest is modified at pH 8.5. Only the α-amino groups are labelled by nicotinic acid.
The ionisation by electrospray and ion spray produces multiple charged ions in the electrical field (e.g. (M+4H)⁴⁺). The existence of multiple charged ions extends the mass range of ESI-MS, because these ions are found in the normal mass range of the spectrometer (usually 150 – 2000 m/z). Due to the fact that always the mass/charge quotient (m/z) is measured in an ion trap or quadrupole instrument, it is possible to calculate the molecular weight of a compound by a simple deconvolution. A single charged ion with the mass of 1000 will be found at the m/z value 1000 as well as a molecule with the mass of 10000 and 10 charges. A normal spectrum of an intact protein does not give a single peak, it produces a series of peaks as depicted in Figure 4-15 A. The intensity decreases from the middle to the outer peaks and the differences between the mass peaks are decreasing towards smaller m/z values. Under the assumption that neighbouring ions differ in the formula by one proton and that the number of additional protons correspond to the charge number of the ion, the molecular weight of a compound can be determined. The m/z scale can now be transformed into a real mass scale as seen in Figure 4-15 B. In C and D the HPLC profile of myoglobin before and after succinylation of amino groups is shown. The succinylation of myoglobin results in an increase in hydrophobicity of the protein at low pH, resulting in an elution at higher % B. The modification of amino groups results in neutralisation of basic residues and leads to a complete unfolding of myoglobin. The combination of these two effects explains the retention on the HPLC column.

4.3.2.3 Derivatisation chemistry III: The effect of nicotinylation on MS/MS

Nicotinylation provides a fast and reliable way to aid the interpretation of mass spectra. As one can easily see from Figure 4-18 the nicotinylation using a mixture of 1-(H₄-nicotinoyloxy)succinimide and 1-(D₄-nicotinoyloxy)succinimide of a succinylated peptide gives a clear visible tag, marking and enhancing all the b-ions. This makes the interpretation of mass spectra much easier, because the b-ion series is clearly distinguishable from the y-ion series and internal series generated due to cleavage at proline peptide bonds. When the b-ion series is clearly identifiable, it is also possible to assign the direction of the sequence, which is not always trivial in non-derivatised spectra. One of the biggest advances using this derivatisation reagent is the fact that the pyridine ring of nicotinic acid functions as a proton acceptor, thus favouring the ionisation capability of the generated fragments, resulting in a significant enhancement of signal intensity.
The spectra of the myoglobin peptide W-Q-Q-V-L-N-V-W-G-B-V-E is shown in Figure 4-17, where B stands for succinylated lysine. In this spectrum, the b-ion series is significantly enhanced in intensity due to the modification of the N-terminus and the presence of tryptophan. The indoyl compound of tryptophan acts as an ideal proton acceptor in the gas phase, thus enhancing the b-ion series. The b^9 ion is not visible in the spectra. This is a frequently observed phenomenon, that there is no abundant ion formation next to glycine residues. This often imposes problems in the interpretation of spectra. Because Gly-Ala is isobaric to Gln and Lys, it is possible to assign the mass difference between two sequence ions in a CID spectrum to Gln/Lys, instead of Gly-Ala. In the spectrum shown in Figure 4-17 A the Gly-Lys difference corresponds to a mass difference of 185, which is close to Trp (186). Due to the modification of lysine by succinic anhydride the mass difference of Gly-sucLys was increased to 286 allowing one to deduce the sequence. In B the MS^3 spectrum is shown from which one can complete the sequence of the peptide.

In Figure 4-18 a peptide from a digest of ribosomal proteins from *Escherichia coli* MC 4100 was analysed. The ribosomes were isolated as described in Hoving et al (1999, submitted for publication), separated on preparative reverse-phase (RP) HPLC system (Machery-Nagel, C_{18}, 250 x 21 mm, Nucleosil 100 – 12 μm; L6220 Intelligent Pump, L-4250 UV-VIS Detector, Merck AG) and fractions were collected. Several fractions have been analysed and the contained peptides were sequenced by the Da/Ha nicotinic acid labelling method. For one fraction (fraction 26) the detailed protein sequence is shown. The fraction was succinylated, purified by RP-HPLC over an RP-8 cartridge, V8 and tryptic digested, Ha/Da-nicotinylated and purified again by RP-HPLC over a RP-8 cartridge. In A a zoom scan 952/956 shows the isotopic distribution due to the modification by 1-(H_2-nicotinoyloxy)succinimide and 1-(D_4-nicotinoyloxy)succinimide. In B the MS/MS spectrum of 952/956 is given. The b_7 ion is not visible in the spectrum, due to its proximity to Arg. The signal of the neutral amino acid Val is almost completely suppressed by Arg, which acts as a proton acceptor due to its high gas phase basicity. The hydroxy amino acids serine and threonine often exhibit loss of water, which appear as ions 18 a.m.u. below the corresponding fragment ion. Such losses are often remote from the cleavage site, thereby resulting in several pairs of ions differing by 18 a.m.u. as can be clearly seen with the b_8^c, b_6^c and b_5^c ion in Figure 4-18. C and D show the MS^3 and MS^4 spectra, allowing the sequence to be read out as N-M-A-G-S-L-V-R.
CHAPTER 4
IDENTIFYING PROTEINS

A: MS/MS (1690/1694)
Myoglobin succinylated, V8 + Trypsin digested; D4/H4 nicotinylated

Figure 4-17 Sequencing of the derivatised myo peptide \( \text{N-Q-Q-V-L-N-V-W-G-B-V-E} \); the peptide was succinylated, purified by RP-HPLC over RP-8 cartridge, V8 and tryptic digested, \( \text{H}_4/\text{D}_4 \)-nicotinylated and purified again by RP-HPLC over RP-8 cartridge A MS/MS spectrum; \( b^{10} \) the succinyl derivative of lysine producing a fragment of 226 a.m.u. The signal corresponding to glycine is not detectable. B MS\(^3\) of the same peptide, allowing to completely define the sequence of the peptide.

B: MS\(^3\) (1690/1694 - 973/977)
Myoglobin succinylated, V8 + Trypsin digested; D4/H4 nicotinylated

Figure 4-18 Sequencing of the derivatised HPLC fraction 26 of E. coli ribosomal protein EC3212 (peptide \( \text{N-M-A-G-S-L-V-R} \)); the fraction was succinylated, purified by RP-HPLC over RP-8 cartridge, V8 and tryptic digested, \( \text{H}_4/\text{D}_4 \)-nicotinylated and purified again by RP-HPLC over RP-8 cartridge A zoom scan 952/956 showing the isotopic distribution due to the modification by \( 1-(\text{H}_4/\text{D}_4)\text{nicotinoyloxy})\text{succinimide} \) and \( 1-(\text{D}_4\text{nicotinoyloxy})\text{succinimide} \), with the characteristic spacing of 4 a.m.u. B MS/MS spectrum of 952/956, note that \( b^2 \) is not visible in the spectrum C and D show the MS\(^3\) and MS\(^4\) spectra, allowing to completely reveal the sequence of the peptide.
A: zoom scan 952/956
HPLC fraction 26 succinylated, V8 + Trypsin digested; D4/H4 nicotinylated

B: MS/MS 952/956
HPLC fraction 26 succinylated, V8 + Trypsin digested; D4/H4 nicotinylated

C: MS^3 (952/956 - 422/426)
HPLC fraction 26 succinylated, V8 + Trypsin digested; D4/H4 nicotinylated

D: MS^4 (952/956 - 422/426 - 351/355)
HPLC fraction 26 succinylated, V8 + Trypsin digested; D4/H4 nicotinylated
**Chapter 4**

**Identifying Proteins**

Figure 4-19  Sequence of 50S RIBOSOMAL PROTEIN L17 (P02416): in bold letters the peptides are given, which were sequenced by the ion labelling method using nicotinic acid derivative

<table>
<thead>
<tr>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>LITLAKTDSV</td>
<td>ANRILAPART</td>
<td>RNEIVAKLF</td>
<td>NELGPRFASR</td>
<td>AGGYTRILKC</td>
<td>GFRAGINAPM</td>
<td>AYIELVDRSB,</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>80</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>127</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRHRKSGRQL</td>
<td>NENSSHQAM</td>
<td>FRRAGISLYR</td>
<td>HEILKTLEPK</td>
<td>AKELHVUEP</td>
<td>KAEAAAE</td>
</tr>
</tbody>
</table>

**4.3.2.4  Derivatisation chemistry IV: The $^{16}$O/$^{18}$O method**

The derivatisation using the isotopic mixture of H$^{16}$O/H$^{18}$O was adapted by Rose and colleagues (Gaskell et al., 1988; Rose et al., 1983; Takao et al., 1991; Whaley and Caprioli, 1991) and rediscovered by Mann and Lehmann (Schnolzer et al., 1996; Shevchenko et al., 1997). Rose carried out a proteolytic digest in buffers prepared with a 50:50 mixture of H$_2^{16}$O/H$_2^{18}$O. Since the protease catalyses incorporation of water, all C-terminal fragments of all generated peptides are characteristically marked by a doublet peak resulting from the difference of two mass units between $^{16}$O and $^{18}$O. This gives the y-ion series a characteristic double peak with a mass difference of two, without any enhancement of the series with a proton acceptor group. The technique has the advantage that no chemical derivatisation of the peptide mixture has to be done, so reducing possible side reactions. It does not increase the abundance of a single ion series, providing a straightforward identification of one type of series. The main drawback lies in the fact, that an unique and error tolerant identification of a double peak needs a very accurate and sensitive mass spectrometer capable of very high resolution such as a Q-TOF (quadrupole time-of-flight) MS. These kinds of instruments are able to deal with very low amounts of material and have a very high mass resolution. In a ‘normal’ triple quadrupole or ion trap instrument, the detector scans the whole mass range by ‘walking’ slowly through the whole mass window in about 3 seconds. A time of flight detector takes snapshots of the whole mass window in one scan with the approximate duration of 1/10 second. This provides the advantage, that the fast time-of-flight detector can obtain 30 spectra in the same time as the slow detector in a triple quadrupole or ion trap instrument. This significantly improves the signal to noise ratio resulting in a much higher resolution. Unfortunately these instruments are twice as expensive as ion trap instruments, which together with the as well expensive H$^{18}$O (1000$\mu$l ~ 1000$\mu$l) limits the applicability of the method. The spectra shown in Figure 4-20 are not as easy to interpret as the spectra shown above.
Figure 4-20 Sequencing of the peptide H-G-T-V-L-T-A-L-G-I-L-K. The peptide was obtained by digesting trypsin in-gel in a 50:50 mixture of H²¹⁵O/H²¹⁸O. A zoom scan of MH²⁺ 690. B MS/MS spectrum of 1378. y-ions show the characteristic double peak with a difference of 2 a.m.u.. b-ions appear as single sharp peaks.
4.4 Discussion

4.4.1 The identification problem

The overall goal of proteomics is to speed up and automate the analysis of large numbers of potential protein targets and to produce data of the highest possible quality. In this respect, it is important to standardise and automate the different techniques and steps used in proteome research to ensure the highest reproducibility and confidence.

To solve the identification problem, two different techniques are available. Firstly, spots of interest can be picked from a gel, concentrated and blotted on PVDF membrane and subsequently analysed by automated Edman degradation chemistry. A second approach would be the analysis of concentrated proteins by mass spectrometry. In the following section, these two approaches will be compared with concerning time, quality, quantity and cost.

4.4.1.1 Mass Spectrometry versus Automated Edman Sequencing: Time

Using an automated Edman sequencer as the Hewlett Packard G1000A, one cycle of Edman degradation chemistry takes about 45 minutes. To sequence a 20mer, which is sufficient to generate primer for subsequent screening approaches, approximately 15 hours of running time are needed. The interpretation of the HPLC read-out is fast and relatively easy and does not need any special knowledge or training.

On the other hand, sequencing by mass spectrometry must be assessed under two conditions. If the peptide or protein of interest is derived from a completely sequenced organism, computer-aided help is provided by algorithm such as SEQUEST, which will be discussed in the following chapter. If the peptide or protein of interest is derived from a non-sequenced organism, the mentioned computer algorithm are useful, but do not prevent the manual and time consuming interpretation of the recorded spectrum. To record a spectrum, approximately 10 minutes per sample are needed, regardless of the length of the peptide. The preparation of a sample for mass spectrometry is comparable to the preparation for Edman sequencing. In the latter case, high attention must be drawn, that the protein of interest is absolutely pure, without any other contamination of protein. Otherwise, a double sequence is obtained in the HPLC read-out, which is almost impossible to interpret. Using mass spectrometry, peptide contamination is a minor problem, because of the nature of experiment. In an ion trap, the ions of interest are se-
lected and kept inside the trap. All other ions, derived from the protein of interest and contamination are discarded. Time consuming in mass spectrometric experiments is only the derivatisation, digestion and desalting steps before the original analysis, which take altogether 2-3 hours of working time. The method can be carried out in parallel on dozens of samples simultaneously and is easily automated.

4.4.1.2 Mass Spectrometry versus Automated Edman Sequencing: Quality

In a standard Edman experiment, only the N-terminal amino acid sequence is revealed. The amount of sequencing data obtainable is in principle only limited by the amount material blotted on the PVDF membrane. There are now a few sequencers on the market (e.g. HP G1000A), which are also able to reveal 2-3 C-terminal amino residues. This might help in the identification of an already known protein, but is far less useful for subsequent genetic screening approach.

The analysis of a protein digest by mass spectrometry routinely gives internal information on the amino acid sequence of the peptides. One advantage of MS is the possibility to obtain additional information about post-translational modifications such as phosphorylation, acetylation or glycosylation. While the acetylation or formylation of an N-terminus of a peptide makes Edman degradation impossible, derivatisation of a peptide on the N-terminus even makes the interpretation of a spectrum easier (see above). Derivatised amino acids or unnatural derivatives can be identified by mass spectrometry as well. One standard draw back of mass spectrometry is the fact that it is only possible to differentiate between the isobaric amino acids Ile and Leu using sectorfield instruments. To differentiate between Gln (Mw 128) and Lys (Mw 128) in a TSQ or IT instrument, a derivatisation step is necessary. This problem is automatically solved in the method described above, which makes the interpretation of spectra easier.

4.4.1.3 Mass Spectrometry versus Automated Edman Sequencing: Quantity

For analysis in an Edman sequencer, 500 femtomoles to 5 pmoles are needed as starting material. The more material provided in the beginning on the PVDF membrane support, the more cycles can be done. It is important to emphasise that the lower the amount of starting material, the higher is the demand on purity. Otherwise, a multiple sequence is obtained which is almost impossible to interpret.
In mass spectrometry, the development of miniaturised electro spray sources has increased the sensitivity to the low femtomole range. With the development of Q-TOF (quadrupole time-of-flight) instruments, even the attomolar range can be reached. Using a triple quadrupole or ion trap instrument, the detector scans the mass window in about 3 seconds. Only these ions are recorded, which pass the detector while the scanning device is active at that position. In a Q-TOF instrument, the detector takes a snapshot of the whole spectrum in 1/10 second. With the same amount of material injected, more spectra can be recorded, resulting in a significant improvement of signal to noise ratio and quality of the spectrum.

Quantity can also be seen under the aspect of automation. While a commercial Edman sequencer processes each sample serially, sequencing by mass spectrometry can be highly parallelised and automated. The desalting and derivatisation steps can be automated and the analysis by mass spectrometry can be fully automated as well by using a commercial autosampler. Only the interpretation of data is providing a bottleneck in this approach. With the SEQUEST program and reasonable computer power, even an online database search is now possible. With a protein or peptide of a non-completely sequenced organism, the derivatisation chemistry described above provides a way to automate the interpretation of unknown peptides and proteins. With the characteristic labelling of the b-ion series described above, the interpretation of the spectra can be done by simple computer algorithm, which will make mass spectrometry even more interesting to non-specialised laboratories.

4.4.1.4 Mass Spectrometry versus Automated Edman Sequencing: COST

The cost of a very sensitive Edman sequencer and an ion trap instrument is quite similar (~300 000 US$) and the maintenance and service costs are comparable. So only the actual running costs have to be compared. The actual costs for sequencing chemicals for one Edman cycle is around 14 CHF (Peter James, personal communication). To obtain the sequence of 20 amino acids 280 CHF have to be taken in account. To sequence by mass spectrometry, only the derivatising labelling reagent, protease (~3 CHF per protein digest) and high purity solvents (~1 CHF) are consumed. Especially in laboratories and industries, which use proteome analysis in a high-throughput approach, e.g. clinical trials, the factors cost, time and automation will become very important in the near future.
4.4.2 Outlook

Protein identification will continue to play an important role in the understanding of biological systems. Because of this, it is anticipated that further developments will be directed towards increasingly affordable, rapid and automated identification technologies. The described ion trap and Q-TOF instruments together with derivatisation chemistry will play a role for the routine and rapid de novo sequencing of peptides. Especially progress in computer-aided interpretation of sequencing spectra will be achieved. One can also imagine that the excellent tolerance to complexity of ESI-MS could be further exploited to routinely allow the identification of more than one protein at a time.

<table>
<thead>
<tr>
<th>Sequence analysis by concerning</th>
<th>MASS SPECTROMETRY</th>
<th>AUTOMATED EDMAN SEQUENCING</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME</td>
<td>Approximately 10 minutes needed to obtain a MS/MS spectrum of a peptide with 20 amino acids; the interpretation of spectra is complicated and needs experience, if not derivatisation chemistry is used</td>
<td>One cycle in an Hewlett Packard G1000A takes 45 minutes; to sequence a 20mer 15 hours running time is needed</td>
</tr>
<tr>
<td>QUALITY</td>
<td>Posttranslational modifications such as glycosylation, phosphorylation and acetylation can be identified</td>
<td>Posttranslational modifications cannot be identified</td>
</tr>
<tr>
<td>COST</td>
<td>Only high purity chemicals as solvents and derivatisation chemicals needed (20mer ~ 4 CHF)</td>
<td>The cost of sequencing chemicals per amino acid is 14 CHF (20mer ~ 280 CHF)</td>
</tr>
<tr>
<td>QUANTITY</td>
<td>N-terminal, internal and C-terminal sequences are routinely obtained by interpreting the spectra; amount: atomol to picomol; automation (high-throughput) possible and commercial available</td>
<td>Only N-terminal sequences can be routinely obtained, blocked or N-terminal modified peptides during preparation are not identified; amount: 500fmol - 1 pmol; high-throughput difficult, not commercial available</td>
</tr>
<tr>
<td>SPECIAL FEATURE</td>
<td>Hydrophobic peptides and proteins are easy to ionise resulting in a strong signal; Ile and Leu are isobaric amino acids and cannot be differentiated by mass spectrometry; to differentiate Gln (Mw 128), Lys (Mw 128) and Glu (Mw 129) a derivatisation step is necessary</td>
<td>Hydrophobic peptides and proteins are washed out with hydrophobic solvents</td>
</tr>
</tbody>
</table>

*Sequencing is automated and can be run 24 hours a day; but not in parallel.**

Table 4-1 Summary of the Mass Spectrometry versus Automated Edman Sequencing
4.5 References


5 Processing Peptide Data

Over the past decade an overwhelming amount of nucleotide and peptide sequence data has been generated and stored in public databases and the growth is now exponential. Automation and the speed of biological analysis has greatly increased, the rate-limiting step is now the processing of the accumulated data. As more data are generated, the larger the sets that have to be processed in order to filter useful information out of the background noise.

Proteome analysis experiments produce a vast amount of data on primary and secondary attributes of a large number of proteins. To correlate these data with the information content of nucleic acid and protein databases, cross-correlation programs are needed to fulfil this task. These programs have to integrate two types of data. First, the experimental data input side, where different information, e.g. molecular weight, pi, processed/un-processed mass data and modifications can be entered. On the database side, different types of database information have to be used and processed, e.g. translation in six reading frames or in silico digests. Subsequently, the two data inputs are cross-correlated with the goal of identifying the best match and an output is generated.

In the following chapter, two programs are presented, where data obtained by mass spectrometry are correlated with nucleic acid and protein sequence information. The identification of a protein is described and the usefulness of these algorithms demonstrated. As an example for the automated detection by SEQUEST, the analysis of Grp94 is shown. An intense Coomassie stained band was obtained after a preparative SDS-PAGE run. The sample was in-gel digested, injected in a HPLC-ion trap LCQ mass spectrometer. The obtained data were processed and the OWL database was searched in a full automated mode by SEQUEST. In addition, an outlook is given of the steps being taken, to create a database with appropriate interrogation tools to identify cell types or bacteria (cell fingerprinting). The results of this chapter will be published in the following articles:

Santella, L., Kyozuka, K., Hoving, S., Münchbach, M., Quadrone, M., James, P. and Carafoli, E. Breakdown of cytoskeletal proteins and calpain action during meiosis resumption by starfish ooocytes. Submitted to Development.

5.1 Peptide Mass Fingerprinting

The idea that the set of (poly)peptide masses obtained by specific enzymatic or chemical cleavage can be a unique fingerprint, allowing a protein to be identified in a database was first put forward in 1977 (Cleveland et al., 1977). The masses of polypeptides generated by in gel proteolysis (estimated from SDS-PAGE gels) were used to identify known viral coat proteins. The mass accuracy obtained was sufficient to allow the identification since the size of the database was so small. However, the idea of using exact peptide masses generated from digested proteins of interest became practical only after the introduction of mass spectrometry.

Measurement with high accuracy at the subpicomolar level of the masses obtained after proteolysis is a prerequisite that cannot be satisfied by gel electrophoresis. With the introduction of new soft ionisation techniques (see chapter 3) such as electrospray ionisation (ESI) and matrix assisted laser desorption and ionisation (MALDI) the analysis of peptides and proteins in biology became feasible. The advent of protein databases and the availability of computers powerful enough to process such large amounts of data fulfilled the second prerequisite for the re-establishment of fingerprinting. In 1993, five independent laboratories published methods and algorithms for database searching using mass spectral data (Henzel et al., 1993; James et al., 1993; Mann et al., 1993; Pappin et al., 1993; Yates et al., 1993). The same general approach was common to all groups. First, a protein is fragmented by a chemical or a protease with high bond specificity (such as CNBr or trypsin) and produces a set of peptides whose molecular masses are the mass profile. Second, a database of mass profiles generated by theoretical fragmentation of a protein database is searched with the experimentally determined profile in a process called peptide mass fingerprinting (PMF). The algorithms used for searching the databases differ significantly, as well as the type of data that can be used in the search. Three of the algorithms (Henzel et al., 1993; Mann et al., 1993; Yates et al., 1993) used a similar scoring scheme, ordering the proteins by decreasing number of matching peptides. The approximate molecular weight range of the intact protein is used as a pre-filter, to eliminate random matches due to the presence of very large proteins in the database. In MOWSE (Pappin et al., 1993) and MassSearch (James et al., 1993); see Figure 5-1) the scoring schemes are based on probability. The algorithm performing MassSearch is best described using the following "ball-in-the-box" analogy: take \(k'\) boxes, each of length \(e\), distribute in the range 0 to 1 and \(n\) random points in the same interval. It can be demonstrated that the probability that \(k\) of the \(k'\) boxes
contain at least one of the \( n \) random numbers is given by \( P = (1 - e^{-nc})^k \). In practice, all peptide weights are normalised to the 0 to 1 interval by dividing them by \((\log(w_{\text{max}}) - \log(w_{\text{min}}))\), where \( w_{\text{max}} \) and \( w_{\text{min}} \) are the highest and lowest peptide weights, respectively. \( K^* \) are the weights used for the search, of which \( k \) match with some of the \( n \) theoretically calculated weights. \( P \) (error interval) gives the probability that such a match happens at random. The calculated probability \( P \) is then converted to a \((-\log_{10}P)\) value (score value) that appears as the score in the matching protein list. The program also provides an indication of the score value below which the match should be considered as not being significant (confidence value). More important than the absolute value of the score is the difference (delta value; score \( \text{ZZZ-YYY} \) in Figure 5-1) between the score of the correct protein and the next non-related protein.

**Experimental** mass list (from MALDI-MS)  
**Calculated** mass list for each database entry

<table>
<thead>
<tr>
<th>868.9</th>
<th>951.5</th>
<th>996.9</th>
<th>1033.0</th>
<th>1059.9</th>
<th>1087.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1109.0</td>
<td>1130.2</td>
<td>1469.3</td>
<td>2257.0</td>
<td>2296.3</td>
<td>...</td>
</tr>
</tbody>
</table>

**MassSearch**  
Searching on SwissProt release 31.  
The sequences are printed in decreasing order of significance.  
Scores lower than XXX are probably not significant.

**Figure 5-1** Peptide Mass Fingerprinting: the MassSearch program (http://cbgr.inf.ethz.ch/MassSearch.html)

The initial reports stressed the wide mass accuracy tolerance of the method since at that stage the protein databases being used contained fewer than 100,000 entries. Nowadays the size is increasing exponentially as a result of the genome sequencing projects, and PMF searches in DNA databases banks has to be performed using a six frame translation of the DNA sequences. Because of the size increase, the confidence level is continually dropping (unless species specific databases are used) and high mass accuracy is required to improve the search accuracy. The introduction of delayed extraction to a reflectron MALDI-TOF led to a routine mass accuracy below 0.05% while maintaining a very high sensitivity (Juhasz et al., 1996; Lennon and Walsh, 1997). Combined with
a relatively good tolerance for low level of contaminants as buffers and salts, this feature made MALDI-TOF instruments the spectrometers of choice for routine fast analysis of unseparated protein digestions. The availability of MALDI targets holding hundreds of digestions and the possibility to automate data acquisition ensures a high sample throughput. The use of PMF to analyse protein mixtures was demonstrated by James et al. (James et al., 1993), where a profile search using data from a digest of casein kinase II holoenzyme (and subunits) was shown. The two subunits were found with scores proportional to the number of their respective peptides used in the search, the scoring being approximately proportional to their molar ratios. One of the main problems of mass mapping lies in assessing the level of confidence that can be placed in the results of protein identification. Studies performed “in silico” (Wise et al., 1997) or on real samples have both shown that an unambiguous identification can be obtained by using two independent sets of experimental data (orthogonal data). This idea was put forward by James et al. (James et al., 1994) and the MassSearch program was modified to allow the use of a second set of masses. The second set can be produced by chemical modification (such as hydrogen-deuterium exchange, methylation or acetylation) of the original digest or by digestion using a different protease. A search using a single digestion (especially with only few peptides) is often not conclusive when performed on a six frame translation of large DNA databases.
5.2 Sequence Tagging

Incomplete protein cleavage often occurs resulting in poor protein sequence coverage for low abundance or hydrophobic proteins and subsequent PMF gives low confidence results. A powerful extension of database searching with MS data is the addition of a small amount of peptide sequence data as a search parameter. Two different approaches have been proposed.

5.2.1 Peptide Fragment Fingerprinting

A peptide can be identified by matching its MS/MS fragmentation spectrum to one predicted from the peptide in the database, which in analogy to PMF can be termed 'peptide fragment fingerprinting'.

The complete manual interpretation of such a spectrum is complex and time consuming (see chapter 4). Since several spectra are collected from a single protein digestion, John Yates’s group at the University of Washington, Seattle, developed a program, SEQUEST, to compare in a fully automated fashion uninterpreted tandem mass spectral data of peptides to theoretically predicted spectra from a protein database (Eng et al., 1994). The program first strips out and averages the MS/MS spectra for each peptide in the MS/MS run and writes the spectrum to an ASCII file, containing a list of peptide masses and ion intensities which is then used to search a database. The database (if DNA, it is first translated in all six reading frames) is searched for all peptides with the appropriate mass ± 3 a.m.u. The MS/MS spectrum of each peptide is then predicted and matched.
to the experimental determined spectrum and the best 500 matches are kept (preliminary score). These peptide spectra are then compared to the experimental one by cross correlation analysis and a list of the top scoring peptides is produced.

5.2.2 Peptide Sequence Tags

Several groups have reported the use of carboxy- and amino-peptidases to generate a ladder of degradation products. The mass difference between the adjacent peaks defines the amino acid being removed and thus the peptide sequence. The development of MALDI-TOF MS has greatly simplified the use of the method: the sample is applied on the target in a series of positions in which a series of dilutions of the enzyme are then added for digestion. The digests obtained on the various positions are then combined to reconstruct the sequence ladder (Patterson and Aebersold, 1995). A different approach is searching the database with the use of peptide sequence tags (Mann and Wilm, 1994). The MS/MS spectrum must be manually interpreted to find a group of ions, which form a series, from which a small sequence, the tag, can be derived. The tags start and end mass, the intact peptide mass and the obtained sequences are used to search the database using the program PepSeach. Recently, our lab developed an algorithm (MassDynSearch) for protein identification using a combination of peptide masses and small associated sequences (tags) generated enzymatically or chemically (Korostensky et al., 1998). The protein of interest is digested and the resultant fragments are subjected to partial exopeptidase degradation. The MALDI-TOF spectra of the digestion before and after the degradation give a list of intact peptide masses each associated with a set of degradation products. The MassDynSearch algorithm uses these 'tagged masses' to search for proteins with similar tagged motifs in protein or DNA databases. The main advantage of this approach is that also less specific proteases (chymotrypsin, elastase, pepsin etc.) can be used in case of proteins harsh to fragment with the specific proteases commonly adopted for PMF. The main drawback is the relatively large amount of material required (high femtomole) and the unpredictable activity of the exopeptidase towards the various peptides making the extraction of ladder sequences from unseparated protein digests very difficult. The chemical procedure (Hoving et al., 1999) uses a water soluble degradation reagent (see chapter 4). In order to reduce sample handling and loss, short ladder sequences are generated via thioacetylation under basic conditions of the N-termini of peptides, which are immobilised on reverse phase material. Subsequent cleavage of the N-termini of the thioacetylated peptides is performed by hydrolysis with acid. The method can be applied to pure peptides or mixture of them, and a prototype machine for its automation has been described (Hoving et al., 1999).
5.3 Application of the Techniques

5.3.1 Peptide Mass Fingerprinting

The heat shock response of \textit{B. japonicum} is a very complex phenomenon. As described in chapter 2, \textit{B. japonicum} posses a variety of control mechanisms and a large number of heat shock proteins. The data presented in chapter 6 were in part generated by PMF. To describe the detection of protein by PMF, the identification of GroEL$_2$ is shown below.

The chaperonins exist as a multigene family consisting of five very similar, but not identical, GroES-like proteins. The \textit{groESL$_1$} operon is controlled by the \textit{\sigma}^{32}\text{-like} \text{RpoH$_1$} RNA polymerase transcription factor. The \textit{groESL$_2$} operon is under control of the \textit{\sigma}^{96} housekeeping RNA polymerase transcription factor. The \textit{groESL$_4$} genes are co-regulated together with the symbiotic nitrogen fixation genes, in that they are activated by the nitrogen fixation regulatory protein NifA at low oxygen conditions and are transcribed \textit{\sigma}^{54} dependent. CIRCE controls the transcription of \textit{groESL$_4$} and \textit{groESL$_5$} (see chapter 2 for references).

To identify GroEL$_2$ in a 2D gel, proteins in the region of interest ~60kDa were picked and processed as described previously. An example of the PMF analysis is shown below, where the output of a search performed with the weight of tryptic peptides of the GroEL$_2$ spot is reported:

Searching on SWISS-PROT release 33.
The sequences are printed in decreasing order of significance.
Scores lower than 77 are probably not significant.

For digester TrpentinCysModified, the fragment weights were:

<table>
<thead>
<tr>
<th>Score</th>
<th>n</th>
<th>k</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>94.7</td>
<td>36</td>
<td>7</td>
<td>P35861;</td>
</tr>
<tr>
<td>80.1</td>
<td>20</td>
<td>5</td>
<td>P28701;</td>
</tr>
<tr>
<td>79.4</td>
<td>33</td>
<td>5</td>
<td>P40047;</td>
</tr>
<tr>
<td>77.1</td>
<td>30</td>
<td>4</td>
<td>P19225;</td>
</tr>
<tr>
<td>77.1</td>
<td>30</td>
<td>4</td>
<td>P19227;</td>
</tr>
<tr>
<td>76.8</td>
<td>14</td>
<td>3</td>
<td>P30735;</td>
</tr>
</tbody>
</table>

The best matches are listed and a score is assigned to each of them. Various parameters are reported and are useful in the evaluation of the match: \textit{n} is the number of theoretical peptides in the
identified protein whose mass is comprised between the lowest and the highest masses used for the search (36) and k (7) is the number of measured peptide masses among that n that matches as mentioned previously, the absolute value in the score is less significant, than the difference in the score between the entry list (delta value: 14.6). Every output is provided with a confidence value (77) below which the match is not considered as reliable. Under AC the database entry code is given. One of the great advantages of the MassSearch program is the ability to integrate in the search the most diverse set of data: the total molecular weight of the protein, data from deuterium exchange experiments, modifications of peptides or multiple digests. As it is shown below, two sets of orthogonal data provide much better results for the identification of the protein, especially, when DNA databases are used as a source. MassSearch screens the entries in genomic databases in both directions in all three reading frames, which increases further the complexity by a factor of six.

Searching on EMBL release 39.
The sequences are printed in decreasing order of significance.
Scores lower than 107 are probably not significant.
For digester TrypsinCysModified, the fragment weights were:
617.600 632.700 682.700 775.900 854.050 1080.400
1232.500 1258.500 1343.400 1560.600 2038.320 2301.100
For digester VRAmmoniumAcetate, the fragment weights were:
1034.200 1344.430 1408.650 1483.540 1938.250

<table>
<thead>
<tr>
<th>Score</th>
<th>n</th>
<th>k</th>
<th>n</th>
<th>k</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>164.9</td>
<td>36</td>
<td>7</td>
<td>12</td>
<td>3</td>
<td>P35861;</td>
</tr>
<tr>
<td>105.9</td>
<td>26</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>P24313;</td>
</tr>
<tr>
<td>105.7</td>
<td>30</td>
<td>4</td>
<td>12</td>
<td>3</td>
<td>Q09708;</td>
</tr>
<tr>
<td>102.6</td>
<td>35</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>P41949;</td>
</tr>
<tr>
<td>100.4</td>
<td>23</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>Q06262;</td>
</tr>
<tr>
<td>98.4</td>
<td>36</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>P36076;</td>
</tr>
</tbody>
</table>

5.3.2 Mining Genomes

Increasing numbers of large-scale genome and EST sequencing projects are releasing vast amounts of sequence data from various types of organisms. In order to deal with this flood of information, John Yates and co-workers developed SEQUEST, a program designed to correlate uninterpreted tandem mass spectra of modified and unmodified peptides, produced under low energy collision conditions with protein and nucleotide derived protein sequences. In a single 30 minutes LC-MS/MS run >200 CID spectra can be generated. The SEQUEST search output on a single MS/MS spectrum can serve to identify a peptide and protein, but multiple MS/MS spectra matching peptides from a single protein give a huge increase in the confidence level. The program is equally
able to deal with multiple proteins in a single digest, and has been used to determine many of pro-
teins in a digest of the whole yeast ribosome (Link et al., 1999).

Figure 5-3 shows an example of protein purification and subsequent analysis by mass spectrometry
as described previously. In A the SEQUEST summary is displayed. It provides a concise overview
of a batch of search results without the necessity of having to look at each individual SEQUEST
output files. It also notes, which proteins are most prevalent in a set of SEQUEST output results.
Each top scoring peptide is displayed along with the input mass calculated from the spectrum
(MassI) and the actual mass of the matched peptide (MassA), correlation scores (Xcorr/DeIcn),
preliminary score (Sp) and rank (RSp), number of ions matched in the preliminary scoring (Ions),
and the protein reference/accession number (Reference) as well as the identified sequence. While
parsing each SEQUEST output file, the occurrence of each protein associated with the top five
ranked peptides is tracked in a 'high score list' at the bottom of A. A score is given to each protein
based on a value of 10 for the top ranked protein, 8 for the second ranked protein, 6 for the third, 4
for the fourth, and 2 for the fifth. Scores are summed for every protein (occurring in the top five
ranking positions) and the top scoring proteins are displayed. By performing this analysis, the
proteins present in a sample are identified since they will score higher because they will be more
prevalent and generally appear at higher rankings, than false positive protein matches.

The most important value is the cross-correlation function, where Xcorr is the normalised correla-
tion score and DeIcn equals 1.0 - normalised correlation score. This provides a measurement of
similarity between the mass-to-charge ratios for the fragment ions predicted by amino acid se-
quencies translated from the nucleotide database and the fragment ions observed in the tandem
mass spectrum. In general, a difference greater than 0.1 between the normalised cross-correlation
functions for the first- and second-ranked search results indicates a successful match between se-
quence and spectrum. Since the preliminary score (Sp) is used to determine the correlation score.
Therefore, a correctly identified peptide usually ranks well for both the correlation score and the
preliminary score.
### Processing Peptide Data

#### Table

<table>
<thead>
<tr>
<th>#</th>
<th>File*</th>
<th>MassI</th>
<th>MassA</th>
<th>Xcorr*</th>
<th>DelCN*</th>
<th>Sp</th>
<th>RSp</th>
<th>Ions</th>
<th>Reference*</th>
<th>Sequence**</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FinnaGECK.0034.0053.1</td>
<td>1306.0</td>
<td>1306.7</td>
<td>2.5991</td>
<td>0.370</td>
<td>149.8</td>
<td>1/11/18</td>
<td>ENPL_CANFA +8</td>
<td>(K)ERPEFLAAGK</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>FinnaGECK.0091.0092.2</td>
<td>2453.0</td>
<td>2452.3</td>
<td>5.2206</td>
<td>0.553</td>
<td>1815.5</td>
<td>1/24/43</td>
<td>ENPL_CANFA +4</td>
<td>(K)TTITTVKLEKSLDASLYKDKIK</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>FinnaGECK.0212.0221.2</td>
<td>3046.0</td>
<td>3046.4</td>
<td>2.6969</td>
<td>0.077</td>
<td>200.8</td>
<td>1/19/100</td>
<td>RRPE_VSYKPM +17</td>
<td>(K)TYLATYGDLEDALQCSSQKSERD</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>FinnaGECK.0171.0188.2</td>
<td>1880.1</td>
<td>1880.9</td>
<td>2.4707</td>
<td>0.024</td>
<td>299.8</td>
<td>12/12/32</td>
<td>PHG2_XENLA</td>
<td>(R)FKCSRTGQGKDNQDGK</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>FinnaGECK.0217.0234.1</td>
<td>1188.0</td>
<td>1187.7</td>
<td>1.7861</td>
<td>0.252</td>
<td>210.9</td>
<td>1/10/20</td>
<td>ENPL_CANFA +4</td>
<td>(K)ZLFVFPAPAFR</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>FinnaGECK.0268.0292.2</td>
<td>2727.0</td>
<td>2727.4</td>
<td>1.7163</td>
<td>0.019</td>
<td>161.9</td>
<td>1/11/42</td>
<td>FS53112</td>
<td>(K)RMCTPQGTVLLEKSNLESDK</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>FinnaGECK.0268.0282.3</td>
<td>4090.0</td>
<td>4090.1</td>
<td>1.6005</td>
<td>0.006</td>
<td>92.4</td>
<td>19</td>
<td>17/140</td>
<td>PS53112</td>
<td>(K)FMCTPEFQKQKLGNLEK</td>
</tr>
<tr>
<td>8</td>
<td>FinnaGECK.0816.0359.1</td>
<td>1485.0</td>
<td>1484.9</td>
<td>1.6767</td>
<td>0.118</td>
<td>161.9</td>
<td>23/8/24</td>
<td>PHG7_XENLA</td>
<td>(R)KIEQAKSLALUK</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>FinnaGECK.0197.0407.1</td>
<td>1595.0</td>
<td>1595.7</td>
<td>2.6804</td>
<td>0.011</td>
<td>227.3</td>
<td>10/9/24</td>
<td>ENPL_CANFA +8</td>
<td>(K)VPFDTHFHMMPK</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>FinnaGECK.0445.0453.2</td>
<td>1797.0</td>
<td>1795.9</td>
<td>5.7340</td>
<td>0.429</td>
<td>2616.0</td>
<td>1/24/30</td>
<td>ENPL_CANFA +4</td>
<td>(K)HEAIQGLOGNAQIR</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>FinnaGECK.0445.0453.3</td>
<td>2680.0</td>
<td>2679.2</td>
<td>2.6907</td>
<td>0.129</td>
<td>251.3</td>
<td>50/18/88</td>
<td>MA2_HUMAN +5</td>
<td>(K)LQKTYSSSEFKAVDPLECABAR</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>FinnaGECK.0486.0493.3</td>
<td>2149.0</td>
<td>2148.3</td>
<td>2.0993</td>
<td>0.094</td>
<td>54.8</td>
<td>145/10/34</td>
<td>CP1I762</td>
<td>(R)THQILQRHRAAQALLLDR</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>FinnaGECK.0486.0493.1</td>
<td>1223.0</td>
<td>1221.5</td>
<td>2.3182</td>
<td>0.039</td>
<td>76.2</td>
<td>265/16/104</td>
<td>HSD268</td>
<td>(K)SRESQKINTSHEKMGREIRFFIKMK</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>FinnaGECK.0522.0526.2</td>
<td>2355.0</td>
<td>2356.2</td>
<td>1.6768</td>
<td>0.078</td>
<td>109.0</td>
<td>9/10/40</td>
<td>HVIP23</td>
<td>(K)TVGSKMAYGDKALKGAIK</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>FinnaGECK.0252.0262.3</td>
<td>3512.0</td>
<td>3512.2</td>
<td>2.3581</td>
<td>0.013</td>
<td>186.1</td>
<td>2/23/120</td>
<td>SKQ9339</td>
<td>(K)NIKAFREYNVEVaVJQVALLLSTYK</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>FinnaGECK.0550.0555.1</td>
<td>1078.0</td>
<td>1078.6</td>
<td>1.3757</td>
<td>0.176</td>
<td>360.7</td>
<td>1/10/18</td>
<td>LU68468</td>
<td>(K)GLTLPSELEK</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>FinnaGECK.0582.0586.2</td>
<td>2289.0</td>
<td>2287.0</td>
<td>2.7793</td>
<td>0.044</td>
<td>154.8</td>
<td>9/10/56</td>
<td>AM5220</td>
<td>(K)NFQCVFLDFQQEQHEVKK</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>FinnaGECK.0590.0614.2</td>
<td>2619.0</td>
<td>2620.5</td>
<td>1.4777</td>
<td>0.081</td>
<td>59.3</td>
<td>52/10/48</td>
<td>SBQ219</td>
<td>(K)EFKSPPHIEQGKNTKPEKTR</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>FinnaGECK.0609.0614.2</td>
<td>3438.0</td>
<td>3439.2</td>
<td>2.2651</td>
<td>0.034</td>
<td>74.1</td>
<td>120/18/132</td>
<td>A27213</td>
<td>(K)DAIQVGMDERLFQKVLLGEKADYDSDYAVYK</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>FinnaGECK.0639.0644.2</td>
<td>2779.0</td>
<td>2779.4</td>
<td>1.2241</td>
<td>0.064</td>
<td>87.7</td>
<td>49/9/46</td>
<td>ST3941</td>
<td>(K)FQEDQAPIEDILTRLQGAPISR</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>FinnaGECK.0639.0644.3</td>
<td>4168.0</td>
<td>4169.1</td>
<td>3.3355</td>
<td>0.041</td>
<td>323.9</td>
<td>25/1544</td>
<td>CASK_HUMAN +3</td>
<td>(O)NHQAMPAEIKTNSVHMLMLGNYFDC MintVSK</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>FinnaGECK.0684.0686.1</td>
<td>804.0</td>
<td>803.4</td>
<td>0.7872</td>
<td>0.007</td>
<td>34.7</td>
<td>238/12/52</td>
<td>SG5267</td>
<td>(K)GAESNWK</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>FinnaGECK.0704.0705.3</td>
<td>699.0</td>
<td>699.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>FinnaGECK.0707.0716.2</td>
<td>1259.0</td>
<td>1267.6</td>
<td>1.0399</td>
<td>0.227</td>
<td>54.2</td>
<td>87/6/22</td>
<td>RI02_TRIKI</td>
<td>(K)DVSPFLGATSK</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>FinnaGECK.0707.0716.3</td>
<td>1903.0</td>
<td>1902.0</td>
<td>3.0993</td>
<td>0.059</td>
<td>67.8</td>
<td>106/15/68</td>
<td>VP6_BTVZ</td>
<td>(K)KVYTVTQSSMAKFRER</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>FinnaGECK.0797.0800.1</td>
<td>858.0</td>
<td>859.5</td>
<td>1.4351</td>
<td>0.090</td>
<td>82.1</td>
<td>53/6/14</td>
<td>HS09383</td>
<td>+1</td>
<td>(K)RASPDVSR</td>
</tr>
</tbody>
</table>

1. ENPL_CANFA
2. RI02_TRIKI
3. ATND_XENLA
4. POST_LXENLA
5. HSU28897
6. HSUG9883
7. ST2091
8. RRPE_MEA3
9. HIVOC3
10. RRPE_VSYKPM
11. A27212
12. SPQ0239
13. CASK_HUMAN
14. VP6_BTVZ
15. FS5312
### Processing Peptide Data

#### Fragment Ion Masses:

<table>
<thead>
<tr>
<th>a-d</th>
<th>a</th>
<th>b</th>
<th>y</th>
<th>v-z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1.01</td>
<td>1785.90</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>102.06</td>
<td>130.05</td>
<td>1656.86</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>231.10</td>
<td>259.09</td>
<td>1527.81</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>360.14</td>
<td>388.14</td>
<td>1398.77</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>431.18</td>
<td>459.17</td>
<td>1327.73</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>544.26</td>
<td>572.26</td>
<td>1214.65</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>672.32</td>
<td>700.12</td>
<td>1096.59</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>795.40</td>
<td>813.40</td>
<td>973.51</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>909.41</td>
<td>928.43</td>
<td>858.48</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>957.45</td>
<td>985.45</td>
<td>801.46</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>1070.54</td>
<td>1098.53</td>
<td>688.37</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>1184.59</td>
<td>1212.57</td>
<td>594.33</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>1235.63</td>
<td>1283.61</td>
<td>503.29</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>1342.65</td>
<td>1370.64</td>
<td>416.26</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>1470.71</td>
<td>1498.70</td>
<td>288.29</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>1583.79</td>
<td>1611.79</td>
<td>175.12</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>1709.89</td>
<td>1737.89</td>
<td>129.02</td>
<td>0</td>
</tr>
</tbody>
</table>

![Fragment Ion Masses Graph](image-url)
G

MRALWVLGLC CVLTFGSVR ADDEVVDGTV VEEDLGKSRE GSRTDDEVVQ REEEAIQLDG LNASQIRESR ESEKFAPQ
EYMRMKLII NLSYKNEKF LRELISNASD ALDKIRLLSL TDENALAGNE ELTVKIEKCK EKNNLHVDVT GQGMTREEFLV
KLNLGIAKSG TSEFINKMT EAQEDQGQSTE LIPQGGVGFY SAFVADKVI VTSKHHNTQ HIWBSDNSF SVIAPDRGNT
LGRRTTLTV LKEEASDYLE LTDKINLTYK YSQFIPINFY WVSSTETVE EPMEEEEAAK EKEDSSDEA AVEEEBEKK
PKTTKVEKTV WDEWLMNDK PIWQRPSEKF EDEEYAFKYP SFKSEDDPM AYIHFATAGGE VTFSKSLFVP TSAPRGLFDE
YGSKKSDYK LYVRVFVTD DPHEMDMPPYL NFVKGVDSD DLPNLSKRET LQQHKLKVI RKLKVREKTD MIKKADKYE
NDTFWKEFGT NIKLQVIEHD SMRTILAKLQ RFQSSSHPSD IITSLDQYVE MKEKQDKYF MAGSRSKEAE SSSPVERLKLK
GKYEVEYLTE PVDECIQLP PFRDGKRFQN VAEEVKVFE DEKTKRSEAA EKEFEPLLN WMKDIALDKD IEKAVVSQRL
TESPCALVAS QYGSNSNER INKQAAYQPG KDISTNNYYAS QKTFRFINPR HPLKMDMLR VKELEDDKTV SDLAVVFLET
ATLRSYLLP DTKAYGDRIE RMLRSLNID PDAKVEEPE ERPEETTED TEDTEQDBEE EMDAGTDDDE QETVKKSTAE
KDEL

Protein coverage: 8.9% of amino acids 9.0% of molecular weight
In an attempt to purify the famous casein kinase from the Golgi apparatus (Bingham et al., 1972) in lactating mammary gland, a purification strategy in combination with an activity test was developed as described by Pinna, James, Münchbach et al. (manuscript in preparation). In the final purification step an intense Coomassie stained band was obtained in a preparative SDS-PAGE run. The band of interest was cut out and processed as previously described. The MS/MS analysis and SEQUEST result is shown in Figure 5-3. The identified protein Grp94 is not a kinase, but is believed to bind to and chaperone the Golgi kinase. This would explain the activity pattern obtained in the purification procedure. It is presumed, that Grp94 was cross-purified with the Golgi kinase.

Grp94 (endoplasmic reticulum paralog of Hsp90) is the sarcoplasmic reticulum paralog of Hsp90 (Cala and Jones, 1994). The sarcoplasmatic reticulum plays a dominant role in cellular Ca$^{2+}$ homeostasis by storing and releasing Ca$^{2+}$. The glycoprotein shows both, Ca$^{2+}$ binding and protein chaperoning properties. Grp94 is also phosphorylated by casein kinase II. It is believed that Grp94 provides a specific protection against Ca$^{2+}$ depletion stress while maintaining the endoplasmic reticulum protein folding function. The amino-terminal region of Grp94 is highly homologous to Hsp90, for which a three-dimensional structure is available (Muresan and Arvan, 1997). This conserved region serves as the peptide-binding site and displays an open/closed conformation for peptide binding. The conserved domain has a nucleotide-binding activity and a weak ATPase activity (Wearsch and Nicchitta, 1997). Peptide binding occurs in an adenine nucleotide-independent manner. It is presumed that accessory proteins regulate Grp94 peptide binding activity through ATP hydrolysis. Grp94 plays also a major role in the trafficking of peptide ligands to the major histocompatibility complex class I molecules that assemble in the ER (Basu and Srivasta, 1999).
5.4 Outlook

5.4.1 Cell fingerprinting

In this section, an outlook is given of the steps being taken, to create a database with appropriate interrogation tools to identify cell types or bacteria. The database will be created using only the information obtained in the HPLC-MS/MS run, without the need of further nucleotide or protein database information. One possible application of the tools for protein identification by this method is the fast and high throughput identification of the originating organism. A typical field of application could be the identification of bacteria responsible for contaminated food or infected drinking water.

In principle three different types of data sets are obtained and serve as parameters in the identification process: the masses acquired during the MS analysis, the isotopic distribution as obtained in a zoom scan and the retention coefficient in the HPLC run.

5.4.2 Sample preparation and processing

*E. coli* and *S. typhimorium* were grown under standard conditions. Aliquots were taken at regular OD intervals between exponential and saturation growth phase. The cells were pelleted, washed several times with an isotonic buffer and subsequently disrupted using a tip sonicator on ice for 5 minutes. Prior to the concentrating and purifying step using a standard concentration gel device, the cells were heated for 10 minutes at 95° C in standard Laemmli buffer. The concentrated protein extract was in-gel digested by trypsin as previously described and extracted for subsequent HPLC-MS/MS analysis. The peptide mixtures were separated by reverse-phase HPLC on a C18 LC Packings Hypersil, 5μm, 300A, 280 x 0.05mm capillary column connected to a Finnigan MAT LCQ ion trap mass spectrometer with a standard electrospray ionisation source as shown in Figure 5-4. The total flow was 3μl/min and the column was washed extensively with solvent A (0.1% (v/v) TFA in H2O) before running a 15min linear gradient from 0-60% of solvent B (80% (v/v) acetonitrile, 0.08% (v/v) TFA). The MS was programmed to accumulate three spectra in a fully automated mode in every 100 a.m.u. window from 500 to 2000 a.m.u: full scan, zoom scan and MS/MS from the most intense ion. Good results could be obtained starting with ca 5x10⁶ cells.
5.4.3 Data extraction from HPLC-MS(//MS) files

The Finnigan LCQ ion trap and TSQ triple quadrupole mass spectrometers store the raw data in a proprietary format that must be extracted into an ASCII file before the data can be processed. A modified version of the LCMS_dta program calculates the charge state of the parent ion used for MS/MS from the zoom scan and then averages the MS/MS files and writes them to an ASCII file.

5.4.4 Creation of a cell fingerprint database with appropriate search tools

This part of the project is being carried out at the Computational Biology Research Group (Prof. G. Gönnet) at ETH Zürich. The data generated are processed using the programming language DARWIN, developed at CBRG at ETH Zürich.

Darwin is a system and programming language suitable for doing computations in biomolecular sciences. Darwin was designed to be a workbench where biochemists could forge tools to explore genetic data quickly and easily. The workbench is a partially interpreted general purpose language containing many built-in routines (Benner et al., 1994) and data struc-
tures (tools from which tools can be built) especially tailored to the needs of bioinformatics community (Gonnet et al., 1992).

Darwin provides a flexible structure to hold a complete or partial genetic database. It is general enough to hold DNA, RNA or amino acid sequences and allows an unlimited amount of annotation information to be kept along-side each entry.

The structure of the database for the cell fingerprinting experiment is relatively simple. The data acquired comprise the retention coefficient from the HPLC run, the mass profile and the isotopic distribution obtained in a zoom scan. These parameters are stored for every sample. A first post-processing of the entries for each specimen is to select which are the masses which are consistently shown in samples. A second post-processing of the entries for all specimens consists of determining which are the masses which are not reliable separators because they appear in too many bacteria. Finally, for a given sample the statistical comparison of the profile against the profile of each entry is developed. The output is a probability indication of how likely it is for each bacterium to be in the sample.

5.4.5 Potential applications

One potential application of cell fingerprinting is in taxonomy. In a future study CBRG will test this using 22 Bifidus bacteria species whose 16S rRNA sequences have been determined and used to create a taxonomic tree. We are currently investigating, if the analysis of the bacterial fingerprints could show whether it is possible to extract biomarkers (sets of masses or certain MS/MS patterns) to define hierarchical relationships between bacteria, genus, species, and strains. If this proves possible the analysis will be extended to compare healthy and diseased cells of the same type to see if specific markers can be found which are characteristic of the disease state. The markers could then be used to program the mass spectrometer to search specifically for the masses entered in a lookup table and to ignore all other masses. Once a mass of interest is found, MS/MS would be carried out on the ion and the spectrum could be compared (on-line) to that of the mass tag. In this way only the tags representative for a disease state would be searched for in a high background of other masses and by adding the MS/MS comparison step, a positive identification could be made.

A potential application of this is the rapid determination of the presence of pathogenic bacteria in samples containing host cells without the need to culture the bacteria. This is especially in view of the many uncultured species recently found by PCR (Bruce et al., 1995).
5.5 References


Cala, S. E. and Jones, L. R. (1994). Grp94 resides within cardiac sarcoplasmic reticulum and is phosphorylated by casein kinase II. J Biol Chem 269, 5926-5931.


6 Functional Proteomics

In the following two articles, examples of functional proteomics are given. As already pointed out in chapter 1, the term functional proteomics emphasises function. It aims to reveal the interdependence of the protein network, their interactions and the all-over behaviour as a result of changes in the cellular environment.

In the following study genomic and proteomic techniques are applied to monitor the behaviour of \textit{B. japonicum} to stress changes caused by heat-shock and various other conditions. A physical map of the ROSE-dependant heat-shock operon was constructed and the transcription-start site determined. In the beginning of the proteomics study, a protein map of the protein complement of \textit{B. japonicum} was constructed by separating and visualising the soluble proteins on two-dimensional gels. To ‘write names on the map for the different locations’, indirect detection by an antibody against DnaK and direct identification methods by Edman and MS/MS sequencing were applied. The raw data were processed using the computer algorithms described in chapter 5 (PMF and PFF) and finally a 2D map of \textit{B. japonicum} was obtained. With this map, it was possible to study the expression behaviour due to changes in the bacterial environment such as heat-shock. Kinetic studies revealed the existence of functional groups of heat-shock proteins, which differ in their heat-shock response.

With this knowledge, it was possible as well to study related species such as Rhizobia, to see if they react in the same way to cellular stress as \textit{B. japonicum}. 

\textit{6-1}
Multiple Small Heat Shock Proteins in Rhizobia

MARTIN MUNCHBACH,1 ANDREAS NOCKER,2 AND FRANZ NARBERHAUS*2

Eidgenössische Technische Hochschule, CH-8092 Zürich, Switzerland

Received 22 July 1998/Accepted 28 October 1998

Seven genes coding for small heat shock proteins (sHsps) have been identified in Bradyrhizobium japonicum. They are organized in five operons that are coordinately regulated by ROSE, a negatively acting DNA element. The deduced sHsps can be divided into two separate classes in which class A proteins show similarity to Escherichia coli HspA and HspB, and class B members display significant similarity to other sHsps from prokaryotes and eukaryotes. Two-dimensional gel electrophoresis and Edman sequencing revealed the presence of at least 12 sHsps in B. japonicum indicating a remarkable abundance of sHsps in this organism. Three additional members of class A and two potentially novel heat shock proteins were identified on the basis of their amino termini. The presence of multiple sHsps was also demonstrated for a variety of Rhizobium and Bradyrhizobium species by immunoblot analysis and two-dimensional gel electrophoresis. An extensive database survey revealed that, in contrast to the rhizobia, other bacteria contain maximally two sHsps whereas many plants have been reported to possess a sHsp superfamily.

Chapter 6 Functional Proteomics
M. Münchbach et al. (J. Bact. 181, No.1, p. 83-90) © American Society for Microbiology, 1999

All organisms so far examined respond to a sudden increase in growth temperature by inducing the synthesis of a number of heat shock proteins (Hsps). Some of these proteins are also important during normal growth. The regulation, structure and function of several Hsps have been studied in great detail. The chaperone machineries GroES/GroEL and DnaK/DnaJ/GroEL, for example, are involved in diverse processes such as protein folding and proteolytic degradation, assembly of protein complexes, and transport of proteins across membranes. Their appearance seems to be highly conserved between prokaryotes and eukaryotes (reviewed in references [7, 40]).

Compared with the highly conserved DnaK and GroEL proteins, sHsps show much less sequence similarity. This protein family is characterized by the following criteria: (i) a molecular mass typically between 12 and 20 kDa; (ii) a conserved central domain, referred to as α-crystallin domain; (iii) formation of large oligomeric complexes, ranging from 150 to 800 kDa; (iv) ATP-independent chaperone activity (9, 27, 51). The latter concept, however, has been challenged by the observation that ATP enhances the molecular chaperone activity of α-crystallin (50).

According to the present model, sHsps bind to denatured proteins accumulated under stress conditions and maintain them in a folding-competent state (15, 32). Most recently, the crystal structure of a sHsp from Methanococcus jannaschii has been solved (29). Twenty-four monomers form a hollow spherical complex with a total of fourteen "windows" that might allow polypeptides to enter the complex.

Not surprisingly, most of the work on sHsps has been conducted with eukaryotic members of this superfamily because they are related to the α-crystallin proteins of the vertebrate eye lens (26). α-Crystallins play a structural role in maintaining lens stability and transparency but notably they are also expressed in nonlenticular tissues, e.g. in heart, muscle and kidney (3). A remarkable abundance of sHsps was reported in heat-stressed plants. Up to 30 different sHsps comprising six different classes are induced after a temperature upshift, depending on the plant species. Each gene family encodes proteins localized in a distinct cellular compartment (51).

Most bacteria appear to have only a small number of heat shock proteins. The completed genome sequences indicate that Mycoplasma genitalium completely lacks any gene coding for sHsps (19). Haemophilus influenzae encodes one and Escherichia coli two sHsps (6, 17). The first hint that Rhizobiales may be an exception in that they possess a larger set of sHsps was provided by a two-dimensional gel analysis by Michiels et al. (34). The authors compared the induction of Hsps in a heat-tolerant and a heat-sensitive Rhizobium strain and observed eight heat-inducible protein spots in extracts from the temperature-sensitive strain. By contrast, the tropical, heat-tolerant strain induced only two sHsps.

In the process of elucidating the complex regulatory network that controls the heat shock response of B. japonicum, the nitrogen-fixing root-nodule symbiont of soybean, we identified six genes encoding sHsps (37, 38). They are organized in four operons that are located in an extended heat-shock gene cluster. Each operon is preceded by a conserved DNA element of approximately 100 bp that is positioned between the transcription start and the start codon of the first gene. This element was designated ROSE (for Repression Of Heat Shock gene Expression), and several lines of evidence suggest that it serves as a binding site for a putative repressor protein under non-heat shock conditions (37).

Here we report on a bacterial sHsp superfamily comprising at least 12 members. The seven B. japonicum sHsps identified so far can be grouped into two distinct classes. We monitored the induction of sHsps under various stress conditions and examined their heat-shock induction by two-dimensional gel electrophoresis. Finally, we provide evidence that the presence of a sHsp superfamily is not restricted to B. japonicum but might be widespread in the Rhizobiales.

MATERIALS AND METHODS
Bacterial strains, plasmids, and growth conditions. B. japonicum 110pse4 was grown aerobically at 25°C in PSY medium (45) supplemented with 0.1% (w/v) arabinose and 100 μg spectinomycin per ml. YEM medium (12) supplemented with 10 mM KNO3 was used for anaerobic B. japonicum cultures. Bradyrhizobium sp. (Parasponia) ANU289 and Bradyrhizobium sp. (Lupinus) ATCC10319 were propagated in PSY medium with 0.1% (w/v) arabinose. TY medium (5) was used to grow Rhizobium leguminosarum bv. viciae 897, R. etli (formerly R. leguminosarum bv. phaseoli) 8002, R. leguminosarum bv. trifolii ATCC14480, Rhizobium sp. strain NGR234, and Sinorhizobium melliloti 2011. E.
RESULTS AND DISCUSSION

Two classes of sHsps in *B. japonicum*. Six genes coding for small Hsps (hspA, B, C, D, E, and F) have recently been identified in a heat shock gene cluster of *B. japonicum* (37, 38). They are organized in four operons together with some additional heat shock genes. Heat-inducible transcription of each operon is mediated by ROSE, a novel regulatory element that consists of approximately 100 bp and precedes the first gene of each operon (37). We identified a putative fifth ROSE-dependent operon by using a ROSE element as a probe in Southern hybridization experiments (data not shown). Two hybridizing fragments, a 5.8 kb BamHI fragment and a 5.6 kb Sall fragment, were subsequently cloned and found to contain the hspH gene region (Fig. 1A). No additional heat shock genes were present up- or downstream of hspH. An amino acid sequence comparison of the deduced small Hsps revealed that they fall into two distinct classes, as indicated in Fig. 1A and shown more precisely in Fig. 2. Class A contains only bacterial proteins, namely the *B. japonicum* proteins HspA, B, D, E, and H, *E. coli* IbpA and IbpB, and *Legionella pneumophila* GspA. It is evident from the alignment that proteins belonging to this class are highly similar to each other throughout their entire length (between 34 to 73 % positional amino acid sequence identity). The similarity is not restricted to the α-crystallin domain but extends into the flanking amino- and carboxy-terminal regions. Class B proteins are much more divergent in length, sequence, and phylogenetic origin. They include prokaryotic as well as eukaryotic members from a wide variety of organisms. The similarity between class A members and class B proteins is rather low (around 20 % amino acid sequence identity). Although the degree of homology within class B is significant (between 30 and 60 % identical amino acids), only the *B. japonicum* proteins HspC and HspF reach the latter, highest score. The identity among the other members is generally between 30 and 35%. Identical amino acids are almost exclusively displayed in the α-crystallin domain, and the flanking regions are highly variable in length and sequence (with the exception of HspC and HspF).

The *B. japonicum* hspH gene is preceded by a typical σ^70^-type promoter and a ROSE element with high sequence similarity to all previously identified ROSE elements (37). With the exception of one nucleotide (a G instead of a C at the ROSE-equivalent position +32), all previously described conserved ROSE nucleotides were conserved in the ROSE element. In particular, the nucleotides in the promoter-distal

FIG. 1. Physical map of five ROSE-dependent heat shock operons of *B. japonicum* and determination of the transcription start site upstream of ROSE5. (A) Schematic representation of the ROSE-dependent operons. The ROSE elements (1–5) are represented by black boxes. Class A and class B small heat shock genes are indicated. No significant open reading frames were identified downstream of hspD and hspH. (B) Primer extension analysis to determine the transcription start site upstream of hspH. The extension product of primer AN18 is shown. The same primer was used for the corresponding sequencing reaction (TCGA).
Multiple Small Heat Shock Proteins in Rhizobia

FIG. 1 Immunoblot analyses of B. japonicum extracts by using anti-E. coli lbpA serum. (A) B. japonicum was grown to mid-exponential phase at 28°C. After a reference sample (28°C) had been taken, the culture was shifted to 45°C and samples were collected at the time points indicated. (B) Similar experiment as in (A) but the cultures were shifted to the temperatures indicated (37, 40 or 43°C), or ethanol (E; 5%), NaCl (N; 0.3M), or H2O2 (H; 0.01%) were added. One extract (anaerobic) originated from a culture grown under anaerobic conditions at 28°C. (C) Fractionation of heat shocked B. japonicum extracts. Normally grown cells (28°C) and heat-shocked cells (45°C) were passed four times through a French pressure cell at 110 Mpa. The soluble (supernatant) and insoluble (pellet) fractions of the heat shocked cells were separated by centrifugation at 12,000 x g for 30 min. The apparent molecular mass of a reference protein (lysozyme; in kDa) is indicated on the right.

half of ROSE are highly conserved (data not shown). Transcription of hspH was heat-inducible and the transcription start site was located at the expected position just upstream of ROSE, as determined by primer extension (Fig. 1B). Thus, all presently known sHsp genes of B. japonicum are coordinate regulated by ROSE, a negatively cis-acting DNA element that precedes each class A gene and presumably serves as a repressor binding site under normal growth conditions (37). The two-dimensional gel analysis Revealed that the degree of induction varied from protein to protein (see below), suggesting that posttranslational mechanisms might contribute to their regulation. Heat-induced expression of class A genes in other organisms (the E. coli lbpAB operon and the L. pneumophila gspA gene) is dependent on a crp-type promoter (1, 2) whereas several class B genes are under negative control. Transcription of the Streptomyces albus hsp18 gene is subject to repression by the OrfY protein at low temperatures (49). Transcriptional repression has also been proposed to control the expression of Clostridium acetobutylicum hsp18, Leuconostoc venus
Induction of sHsps in *B. japonicum*. A so far undefined set of *B. japonicum* sHsps was recognized by an antiserum raised against the 15 N-terminal amino acids of *E. coli* hspA which is similar to the *B. japonicum* class A proteins (2, 39). Three cross-reacting bands were detected in extracts from heat-shocked *B. japonicum* cells, and the fastest migrating band was absent in a hspBC mutant (*B. japonicum* 5069) indicating that the antiserum specifically recognized the HspB protein (39). Immunoblots of two-dimensional gels revealed that the serum indeed recognizes several class A proteins (HspB, HspD, HspH, and spots 4, 8, and 10; data not shown, compare with Fig. 4). We monitored the kinetics of sHsp induction in *B. japonicum* by using this antiserum. Extracts from cells harvested before and at different time points after a heat shock from 28°C to 43°C were analyzed (Fig. 3A). The first faint signal was observed at 5 min after the heat shock. The accumulation of sHsps continued until the level reached a maximum approximately 60 min after the temperature upshift. This elevated level was maintained for at least another hour. In a separate experiment, we determined how a shift to various temperatures affected the induction of sHsps. The amount of sHsps increased proportionally to the severity of the shift (Fig. 3B). A shock from 28 to 37°C was sufficient to induce the complete set of immunodetectable sHsps, but a shift to 40°C and in particular to 43°C was much more efficient. Next, we analyzed whether other stress conditions could induce the synthesis of class A sHsps. Cultures grown at 28°C did not induce sHsps during the onset of, or in, stationary phase (data not shown). Extracts from bacteroids that had been isolated from soybean root nodules did not contain detectable amounts of sHsps (data not shown). Neither continuous growth under anaerobic conditions, nor a shift of aerobically grown cultures to high salt (0.3M NaCl) conditions or to a highly oxidizing environment (0.001% H2O2) elicited a significant response (Fig. 3B). However, the addition of ethanol (5%) to a culture led to the production of sHsps, albeit to a much lesser extent than a heat shock. This result suggests that both, a temperature shift and an ethanol shock, trigger a signal that is finally transduced to induce the synthesis of sHsps in *B. japonicum*.

Many studies indicate that certain sHsps in animals, plants, and bacteria are regulated by a variety of environmental and developmental cues. Developmental synthesis of sHsps in eukaryotes is often tissue-specific in contrast to the coordinate heat-shock induction of sHsps in almost all tissues. Constitutive, but low expression of Hsp27, a mammalian sHsp, was observed in different cell types. This protein plays a role in regulating the dynamics of actin filaments and probably confers stability to actin fibers [reviewed in

**FIG. 4.** Two-dimensional gel electrophoresis of *B. japonicum* extracts. Crude extracts of cells grown at 28°C (A) or of cells shifted from 28 to 43°C for 30 min (B) were separated and stained with Coomassie blue. DnaK, a set of GroEL proteins, and sHsps are circled. The positions numbered with 1 to 11 in (A) correspond to the proteins listed in Table 1. HspB, C, D, E, and H that were identified by amino-terminal sequencing are labeled in (B) with B, C, D, E, and H, respectively. Induction of sHsps in *B. japonicum* WT (C) and the hspBdegP mutant *B. japonicum* 5069 (D) (34). Relevant sections of two-dimensional Coomassie-blue stained gels are shown. The positions of HspB and HspC which are present in the wild type but missing in the mutant are indicated by an arrow. Other sHsps in panels C and D are circled.
Multiple Small Heat Shock Proteins in Rhizobia

Reference (3). Expression of plant sHsps during pollen development, seed and fruit maturation has been reported. Again, only a subset of the sHsps reacts to the developmental signals, and their expression is temporally and spatially controlled (31). A number of bacterial sHsps can also be induced by developmental signals although heat shock often is the major elicitor. L. pneumophila GspA is expressed during intracellular infection of macrophages and mycobacterial Hsp16 might also be induced in response to stresses encountered during an infection process (1, 52). The Bacillus subtilis CodM protein is developmentally induced during sporulation and Stigmatella aurantiaca SP21 is synthesized during sporulation and fruiting body formation (24, 25). Induction of C. acetobutylicum Hsp18 was demonstrated during a metabolic shift from acid to solvent production (41, 47). By contrast, our investigation indicates that at least the immunodetectable B. japonicum sHsps are classical heat stress proteins.

Small Hsps aggregate after heat shock in vivo. Extracts of heat shocked B. japonicum cells were separated into a soluble and insoluble fraction. The immunodetectable sHps were almost exclusively found in the pellet fraction (Fig. 3C) indicating that they form insoluble aggregates after heat shock. Whether these aggregates consist only of sHsps (homo- or heterooligomers), or whether substrate proteins are bound to the sHsps cannot be determined at present.

Identification of B. japonicum sHsps by two-dimensional gel analysis. The presence of at least seven genes coding for sHsps in B. japonicum prompted us to investigate the induction of such proteins by comparative two-dimensional gel electrophoresis (Fig. 4). The positions of DnaK and GroEL are indicated for comparison (Fig. 4A and B). Note that B. japonicum contains five groESL operons and that the GroEL spot represents a composite of several GroEL proteins ([16]; M. Münchbach and H. M. Fischer, unpublished results). At least 11 small proteins were reproducibly upregulated after a heat shock and visible on Coomassie-stained two-dimensional gels. GroESL, HspB, C, D, E, and H were identified by N-terminal sequencing of the collected protein spots from several gels (Fig. 4B). A comparison of the sHsp pattern after heat shock in the wild type and the suppBCdegP mutant 5069 confirmed the identity of the HspB and HspC spots because they were in fact missing in the mutant (Fig. 4C and D). HspA and HspF could not be identified. HspA may not be detectable due to a cathodic drift in the first dimension (calculated isoelectric point of 8.42). The amount of HspF is probably too low to be detectable because HspE, the product of the first gene of the hspEForG operon, is also barely visible. The amino termini of proteins 4, 8, and 10 (MRTYDLTP, MRTYDFLP, and MRSYDFSPLWRSTXR, respectively; compare with Fig. 2) indicated that B. japonicum contains at least three additional class A sHsps whose structural genes and regulatory elements have yet to be identified. The amino-terminal sequence of two proteins (ALYFHVFL and AG1VLQKL for spots 2 and 5 in Fig. 4, respectively) did not show similarity to class A or class B proteins or any other proteins in the databases which suggests that there might be additional sHsp classes in B. japonicum. In summary, we predict that B. japonicum contains a total of at least 12 sHsps.

A set of sHsps is present in other rhizobia. In order to test whether a superfamily of sHsps is present in other rhizobial species, we screened a variety of Bradyrhizobium and Rhizobium strains by immunoblot analysis using the anti-E. coli HspA serum. Heat-induction of one or several bands was observed in each case indicating that all species tested possess class A-type sHsps (data not shown). To monitor the heat-induced proteins more accurately, we performed two-dimensional gel electrophoresis of extracts from six rhizobial
### TABLE 1. Number and classification of small heat shock proteins in various organisms

<table>
<thead>
<tr>
<th>Organism</th>
<th>Class A</th>
<th>Class B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bradyrhizobium japonicum</td>
<td>$\geq 8^*$</td>
<td>$\geq 2$</td>
<td>$\geq 12^*$</td>
</tr>
<tr>
<td>Bradyrhizobium sp. (Parasponia)</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Bradyrhizobium sp. (Lupinus)</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Rhizobium etli CNPAF512</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Rhizobium etli 8002</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Rhizobium sp. strain NGR234</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Rhizobium leguminosarum bv. trifolii</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Rhizobium leguminosarum bv. viciae</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Sinorhizobium meliloti</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Rhizobium tropici CIAT899</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
</tr>
<tr>
<td>Mycoplasma genitalium$^1$</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Haemophilus influenzae$^2$</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Helicobacter pylori$^2$</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Borrelia burgdorferi$^2$</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Treponema pallidum$^2$</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Escherichia coli$^2$</td>
<td>2</td>
<td>–</td>
<td>2</td>
</tr>
<tr>
<td>Methanococcus jannaschii$^2$</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Aquifex aeolicus$^2$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Archeoglobus fulgidus$^2$</td>
<td>2</td>
<td>–</td>
<td>2</td>
</tr>
<tr>
<td>Rickettsia rubida$^2$</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis$^2$</td>
<td>–</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae$^2$</td>
<td>2</td>
<td>–</td>
<td>2</td>
</tr>
<tr>
<td>Rhodobacter capsulatus</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Legionella pneumophila</td>
<td>1</td>
<td>n.d.</td>
<td>1</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>1</td>
<td>n.d.</td>
<td>1</td>
</tr>
<tr>
<td>Synecocystis sp. strain PCC6803</td>
<td>n.d.</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Stigmatella aurantiaca</td>
<td>n.d.</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>n.d.</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Streptomyces albus</td>
<td>n.d.</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>–</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>Glycine max (soybean)</td>
<td>n.d.</td>
<td>$\geq 10$</td>
<td>$\geq 10$</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>n.d.</td>
<td>$\geq 12$</td>
<td>$\geq 12$</td>
</tr>
<tr>
<td>Oryza sativa (rice)</td>
<td>n.d.</td>
<td>$\geq 11$</td>
<td>$\geq 11$</td>
</tr>
</tbody>
</table>

**Footnotes:**
- $^*$ based on DNA sequence analysis and amino terminal sequencing of isolated proteins
- $^\dagger$ as determined by two-dimensional gel analysis
- $^\ddagger$ not determined or not known because the genome has not been sequenced
- $^\dagger\dagger$ as determined by one-dimensional gel electrophoresis and immunoblot (data not shown)
- $^1$ complete genome sequence available
- $^2$ YocM and CotM are defined as $\alpha$-crystallin-type proteins in the corresponding SubList database. However, YdfT has also some similarity to class B proteins (around 20% overall sequence identity) and might fall into this class.
- $^3$ human and animal sHsps appear to fall into a separate class distinct from class A and B

* species. Between 3 and 10 potential sHsps were observed in each strain (Fig. 5; see Table 1). In summary, we conclude that the existence of a sHsp family is not restricted to B. japonicum but occurs in many rhizobial species.

The presence of multiple sHsps in a bacterium is a rather uncommon feature. A literature and database survey, including the 35 microbial genomes that are completed or currently being sequenced, revealed that bacteria other than rhizobia encode either no or maximally two sHsps (Table 1). For example, no $\alpha$-crystallin-like protein was found in the genomes of the pathogens Mycoplasma genitalium, Haemophilus influenzae, Helicobacter pylori, and Borrelia burgdorferi. The available sequence of Rhodobacter capsulatus, an $\alpha$-proteobacterium and close relative of rhizobia, also did not reveal any sHsp. One or two sHsps are encoded in the genome of a number of eubacteria and archaeaebacteria and in yeast. Interestingly, if one of these organisms contains two sHsps, they always belong to the same class.

The existence of a sHsp superfamily comprising defined classes is well-established in plants (Table 1). For example, the sequences of ten soybean (Glycine max) sHsps are deposited in the public databases. They clearly fall into class B but have been further subdivided in different subfamilies. Six groups were classified: two classes (class I and II) localized to the cytosol, and one class each localized to the chloroplast, endoplasmic reticulum, mitochondrion and membrane compartment (51). The homology between individual members of these classes is restricted to only a few amino acids in the $\alpha$-crystallin domain. An phylogenetic analysis suggested that the abundance of plant sHsps arose from an ancient gene duplication or amplification more than 150 million years ago that was followed by sequence divergence (51). A similar gene multiplication event with subsequent diversification might have occurred in B. japonicum giving rise to the unusual broad spectrum of bacterial sHsps. The localization of six B. japonicum genes (hspA to hspF) encoding sHsps in a heat shock gene cluster probably supports this assumption. Five human sHsps have been described (13). The ongoing genome sequencing projects will reveal whether sHsp superfamilies are common in mammals.
It is unclear why the rhizobia analyzed in this work contain multiple sHsps whereas most other organisms do not. The relative abundance of rhizobial sHsps after a heat shock certainly implies an important cellular function offering an advantage in their natural environment. Short periods of intense sunlight, for example, might cause protein damage. When chaperones become temporarily overloaded with potential substrates, sHsps might play an important role as buffer for otherwise aggregation-prone enzymes. In agreement with a recent model (15, 32), one can imagine that this reservoir of folding-competent proteins will later be refolded by the cellular chaperone machineries under conditions when their capacity becomes available again. For example, the tropical Rhizobium strain that is adapted to high temperatures apparently does not require multiple sHsps because it contains only two small heat-inducible proteins (34). The reason for the heat-tolerance of this strain is unknown. Bacteria which thrive as mammalian pathogens live in an environment with more or less constant temperatures and may be able to cope without a sophisticated heat shock response. Their lifestyle is reflected by a comparatively small number of sHsp genes in their genome.

ACKNOWLEDGMENTS

We are grateful to Hanke Hennecke and Peter James for generous support, continuous interest in our work, and helpful comments on the manuscript. Hans-Martin Fischer and Evelyn Bauer are acknowledged for providing B. japonicum extracts. Rhizobium strains were obtained from Michael Güttlert. We thank Alan Easton for the generous gift of antisera, and Wolfgang Weigler for performing the experiment shown in Fig. 3A. This study was supported by grants from the Swiss National Foundation for Scientific Research and the Swiss Federal Institute of Technology, Zürich.

REFERENCES


Proteome Analysis of Heat Shock Protein Expression in Bradyrhizobium japonicum

Martin Münchbach1, Paola Dainese1, Werner Staudenmann1, Franz Narberhaus2, and Peter James1

1Protein Chemistry Laboratory, Swiss Federal Institute of Technology, Zürich, Switzerland; 2Institute of Microbiology, Swiss Federal Institute of Technology, Zürich, Switzerland.

A set of nineteen heat shock proteins (Hsps) was observed by subtractive two-dimensional gel electrophoresis to be induced when Bradyrhizobium japonicum, the nitrogen-fixing root-nodule symbiont of soybean, was temperature up-shifted from 28°C to 43°C. Up-regulated protein spots were excised from multiple two-dimensional gels. The proteins were concentrated using a funnel-gel device before being blotted onto PVDF membranes for digestion with trypsin prior to MS and MS/MS analysis or for Edman sequence determination. Five proteins in the range of 8-20 kDa were identified as the small heat shock proteins (sHsps). Two other low-molecular weight proteins corresponded to GroELs and GroESs, and five novel proteins were found. Four or five 10-30 kDa proteins were identified as GroELs, GroESs, and GroELs and DnaK. An analysis of the heat shock induction of DnaK, one of the most strongly induced GroEL proteins and six of the sHsps revealed that the proteins could be placed into four distinct regulatory groups based on the kinetics of protein appearance.

Keywords: heat shock; kinetics; mass spectrometry; multiple-gene families; proteome

Virtually all cells, both eukaryotic and prokaryotic, respond to environmental stress by an increase in the synthesis of a set of proteins termed heat shock proteins. There are six conserved classes of heat shock proteins that have chaperone activity: Hsp100, Hsp90, Hsp60, Hsp40 and the small heat shock proteins (sHsps). The Hsp100/Clp proteins are a family with a wide variety of functions, such as increased tolerance to high temperatures through ATP-dependent disaggregation and unfolding for degradation, promotion of proteolysis of specific cellular substrates and regulation of transcription [1]. The Hsp90 family is also part of the powerful network of chaperones to prevent incorrect interactions of proteins. They play a fundamental role in cellular processes such as hormone signalling and cell cycle control. The regulation of key signalling molecules has made the Hsp90 machinery a promising new drug target [2,3]. The Hsp70 and the Hsp90 family typified by the bacterial DnaK and the bacterial GroEL respectively have been extensively studied. Their role in protein folding, degradation and in the assembly of large protein complexes is well documented [4]. The crystal structures of Hsp90 [5], the GroEL/GroES complex [6], the Hsp40 J-domain [7] and of GroE with the ATPase domain of DnaK [8] have been solved.

In contrast, much less is known about the small heat shock proteins though recently the crystal structure of a sHsp from a hyperthermophilic organism, Methanococcus jannaschii, has been solved [9]. They form a much more diverse family though they have some properties in common such as: a conserved α-crystallin homology domain; a molecular weight between 10 and 30 kDa although they tend to form large oligomeric assemblies. The sHsps have a wide range of cellular functions, including the ability of cells which have a thermo-tolerance to maintain folding of microfilaments and being able to act as molecular chaperones in vitro, sHsps do this by forming stable complexes with folding intermediates of their protein substrates [10, 11]. They enhance the survival of eukaryotic cells exposed to oxidative stress by decreasing the intracellular level of reactive oxygen species in a glutathione dependent way and they seem to be a novel regulator that interferes with programmed cell death [12]. Their main function appears to be the prevention of the accumulation of unfolded protein intermediates during stress periods [13, 14].

In contrast to the extensive eukaryotic gene families (e.g. 20 Hsp70 homologous in yeast [15]), most bacteria appear to have a limited set of one or two heat shock proteins per class. Unusually, rhizobial bacteria appear to have a set of multi-gene heat shock protein families like the GroE/S [16] and small heat shock proteins [17, 18]. Bradyrhizobium japonicum, the nitrogen-fixing root-nodule symbiont of soybean, possesses at least five GroEL and seven sHsp genes. This organism has proven to be a valuable prokaryotic model system to study both the regulation and function of paralogous gene families. The roles of some of the GroE/S proteins are partially understood [16, 19], but little is known about the sHsps. B. japonicum and other rhizobial species are clearly unique among eukaryotes because they induce sHsps in a plant-like fashion. At least 12 sHsps are induced upon heat shock. These proteins can be divided into two separate classes: class A consisting of proteins that show similarity to Escherichia coli hspA and hspB and class B, whose members display significant similarity to other sHsps from prokaryotes and eukaryotes [18]. Whether the individual members of these two classes have distinct functions in vivo is the subject of ongoing...
research in our laboratories. In an attempt to assign functional roles to the multiple heat shock genes we undertook a study of how B. japonicum responds to a temperature shift from 28°C to 43°C by analysing the changes in protein expression using two-dimensional gel electrophoresis. We present data identifying nineteen heat shock induced proteins and an analysis of the kinetics of induction of eleven of these. The sHsps induction kinetics can be distinguished from that of the GroEL and can be subdivided into two distinct subgroups, despite possessing an almost identical regulatory motif [20] at the beginning of their respective operons.

MATERIALS AND METHODS

Materials

Acrylamide, N,N'-methylenebisacrylamide and carrier ampholytes for two-dimensional electrophoresis were purchased from BDH (Poole, England); CHAPS was from Sigma (Buchs, Switzerland); Coomassie Brilliant Blue (Serva Blue G 250) was from Serva (Heidelberg, Germany); Immobiline™ strips were from Pharmacia (Uppsala, Sweden). All other reagents for 2D-PAGE were the highest purity grade available from Fluka (Buchs, Switzerland). Fluorotrans PVDF membrane was obtained from PALL (Mutenz, Switzerland) and β octyiglucopyranoside was from Pierce (Rockford, IL, U.S.A). Sequencing grade modified trypsin was purchased from Promega (Zürich, Switzerland) and DNase and RNase were from Boehringer (Mannheim, Germany). All HPLC solvents used were from Riedel deHaen (Seelze, Germany).

Bacterial strains, growth conditions and cell extraction

B. japonicum 110pG4 was grown aerobically at 28°C in PSY [21] supplemented with 0.1% (w/v) arabinose and 100 µg spectinomycin per ml. The culture was harvested at A600nm=1.2 by centrifugation (6000xg, 10 min, 4°C) and washed three times with 0.9 M NaCl. The cells were then resuspended in 50 mM Tris/HCl, pH 7.0 (100 mg wet weight/100 µl) and ruptured by sonication. The solution was treated with DNase I (50 µg/ml) and RNase A (10 µg/ml) at 37°C for 30 minutes. The proteins were solubilized in 1% w/v SDS, 150 mM DTT, before equilibration in 8 M Urea, 2% CHAPS, 10 mM DTT and 0.8% carrier ampholytes pH 4-12. Cell debris was removed by centrifugation (12000xg, 30 min, 4°C).

Two-dimensional gel electrophoresis

Two-dimensional electrophoresis was carried out as described by O'Farrel [22] but using immobilised gradient gels in the first dimension [23]. Immobiline™ strips (Pharmacia, Uppsala, Sweden) were rehydrated overnight in 8 M urea, 2% CHAPS, 10 mM DTT, and 0.8% ampholytes pH 4-8, 100 mg of protein was loaded onto each Immobiline™ strip in a gel rehydration cassette [24]. The first dimension Immobiline™ strips were run in batches of 20 on a Multiphor II system (Pharmacia, Uppsala, Sweden). The samples were focused at 300 V for three hours, then the voltage was ramped up to 3500 V over one hour. The final V x h was fixed at 80000. The strips were then equilibrated for 20 min in 50 mM Tris/HCl pH 6.8, 6 M urea, 25% glycerol, 0.2% SDS, 65 mM DTT, then incubated for 5 min in 50 mM Tris/HCl pH 6.8, 6 M urea, 25% glycerol, 0.2% SDS, 65 mM iodoacetamide before being transferred onto 12% constant polyacrylamide gels. The second dimension gels were run in batches of 20 using an ISO-DALT apparatus (Hoefer, San Francisco, CA, U.S.A.). Gels were stained with Coomassie Blue as described by Schägger and von Jagow [25] and the wet gels were scanned in a Personal Densitometer (Molecular Dynamics, Sunnyvale, CA, U.S.A.).

Kinetic analysis

Two litres of bacterial culture were grown at 28°C until the absorbance A 600nm reached 1.2 at which time the culture was divided into 200 ml aliquots for heat shock treatment at 43°C for defined periods. After heat shock the cells were harvested by centrifugation and then frozen immediately in liquid nitrogen. Preliminary analyses were carried out after 5, 15 and 30 minutes of heat shock. For the detailed kinetic analysis four batches of bacterial culture were grown and samples taken at 0, 2, 4, 6, 8, 10, 12, 14, 20, and 30 minutes. To ensure reproducible running conditions in both dimensions for all gels, the samples were run in batches of 20 using a Multiphor II system (Pharmacia, Uppsala, Sweden) and an ISO-DALT apparatus (Hoefer, San Francisco, CA, U.S.A.). Eighty two-dimensional gels were run (samples from four separate heat shock experiments), stained with Coomassie Blue under identical conditions before being imaged with a laser scanning densitometer and analysed using the 2-D™ gel analysis program from Protein Databases Incorporated (New York, NY, U.S.A.).

The images from each time point were used to create a single master image. A 'super-master' image was then created for the entire time set using the master images. The integrated optical density of the spots was calculated as a percentage of the total density of the 'super-master' gel and the values exported into the statistical data analysis package Cricket Graph for presentation. For each time-point, a master image was constructed, and then the integrated optical density for each spot determined as a percentage of the total stained protein. The data was plotted using Cricket Graphic III (Cricket Software, Malvern, PA, U.S.A.).

Protein elution, concentration and electrophor transfer onto PVDF membrane

Spots cut out of two-dimensional gels were washed three times for 30 minutes in excess double distilled water and subsequently incubated overnight in the minimum volume of Laemmli solubilisation buffer. Protein spots were concentrated to single sharp bands using a funnel-shaped gel electrophoresis device [26] and then electroblotted onto PVDF membranes in a semi-dry apparatus (Hoefer, San Francisco, CA, U.S.A.) in a buffer containing 50 mM Tris/HCl, 192 mM Glycine, 0.02% w/v SDS, 10% v/v methanol, 2 mM DTT for 1 hour at approximately 1.2 mA/cm². Proteins were visualised on the membranes by a 5 minute incubation in 0.1% w/v Serva Blue R in 50% v/v methanol, followed by destaining in 70% v/v methanol for 5-10 min.

Edman sequencing and protein digestion

Proteins blotted onto PVDF were analysed on a Hewlett Packard G1000A by Edman sequencing. For each protein, one third of the membrane was used for N-terminal analysis while the remaining was subjected to on-blot digestion with trypsin. The PVDF membrane was cut in very small pieces and equilibrated for 1 hour at room temperature in 10 µl of 1% β-octyglycophoranside, 6 M urea, 25% glycerol, 0.2% SDS, 65 mM iodoacetamide.
100 mM ammonium bicarbonate (pH 7.8) and digested for 15 h at room temperature using modified Trypsin (1 μg/μl). Adding 1 μl of 2% v/v TFA to the sample stopped digestion. The supernatant was collected and analysed by MALDI-TOF-MS on a Voyager Elite (Perseptive Biosystems, Framingham, MA, U.S.A.) as previously described [27].

Automated peptide fragmentation by collision-induced dissociation

Digestion mixtures were separated by reverse-phase HPLC on a capillary column (C18, 5 μm, 300 Å, 280 x 0.05 mm) from LC Packings International (Zürich, Switzerland) directly connected to a Finnigan MAT (San Jose, CA, U.S.A.) TSQ 700 triple quadrupole mass spectrometer equipped with an electrospray ionisation source using a coaxial flow of 1.5 μl/min methoxyethanol as a sheath liquid. The total flow was 3 μl/min and the column was washed extensively with solvent A (0.1% v/v TFA in H2O) before running a 60 min linear gradient from 0 to 70% solvent B (80% v/v acetonitrile, 0.08% TFA). For analysis a modified version of Autofrag, which automates the collection of CID fragmentation spectra from unknown samples was used as described [27]. The automated procedure controlled by Autofrag allows both parent ion mass measurement for protein mass fingerprinting and sequence analysis by fragmentation for peptide mass fingerprinting to be carried out during the same HPLC run. The peptide of interest is isolated from other coeluting peptides using the first mass filter Q1, before being accelerated into the collision gas chamber in the second stage Q2. The fragments generated by multiple collisions are analysed in the third mass-scanning device Q3. Masses of known contaminants as trypsin or keratin fragments were excluded and not used for carrying out MS/MS analysis.

RESULTS

Two-dimensional gel analysis of heat shock protein induction in *Bradyrhizobium japonicum*

In previous studies [18] the kinetics of sHsp induction in *B. japonicum* was monitored on one-dimensional SDS-PAGE using a polyclonal antiserum raised against the 15 N-terminal amino acids.

**Fig. 1. Heat shock protein induction varies with time.** The low molecular weight region of two-dimensional gels of protein extracts from *B. japonicum* cells after different times of exposition to a heat shock at 43°C. No further changes were observed between 30 and 60 minutes.

(A) *B. japonicum* grown at 28°C

(B) *B. japonicum* after 5 min heat shock at 43°C

(C) *B. japonicum* after 15 min heat shock at 43°C

(D) *B. japonicum* after 30 min heat shock at 43°C
of *Escherichia coli* IbpA, which is similar to the N-terminus of class A sHsps. It was found that the induction of heat shock is most effective at 43°C and takes place in a time window from 5 to 30 minutes. To overcome the drawback of the poor resolving power of a one-dimensional gel and the possible cross-reactivity of a polyclonal antibody a two-dimensional gel analysis was carried out and single spots were identified using proteomic tools. Preliminary experiments using two-dimensional electrophoresis monitored the extent of the induction of the different class members of small heat shock proteins in *B. japonicum* after a temperature shift from 28°C to 43°C. The induction was found to be time dependant. Samples were taken at various intervals between 5 and 30 minutes and subsequently analysed by two-dimensional gel electrophoresis (Figure 1). A set of thirteen small proteins was observed to be induced by the temperature shift. In order to identify the induced proteins, a batch of 20 gels was run; four gels were loaded with control samples (i.e. from bacteria grown at 28°C) and 16 gels with cell extracts from *B. japonicum* exposed to 43°C for 30 minutes. After fixation and staining, the gels were scanned with a laser densitometer and the resulting images analysed. Nineteen protein spots were observed to be induced, twelve de novo and seven others were clearly upregulated (Figure 2). The average protein was present at around 0.5-3 pmol/spot. Edman analysis

![Figure 2. Identification of heat shock proteins. Two-dimensional gel electrophoresis mapping of *B. japonicum* in non-heat shocked and heat shocked state (30 minutes at 43°C). A circle indicates the proteins that are induced during the heat shock. These proteins were identified by MS, MS/MS analysis of their tryptic digests and Edman degradation (see Table 1).](image)

![Figure 3. MS analysis of the tryptic digestion of spot 11. MALDI-TOF-MS spectrum of the tryptic digestion of protein spot 11. On the basis of the peptide masses observed, the protein could be identified as GroES.](image)
Table 1. Identification of the 19 heat shock proteins labelled in Fig 2. Proteins induced by a 30 minutes heat shock at 43°C. The proteins are numbered according to Figure 2. Database searches with the N-terminal sequences were performed using the FASTA and TFASTA [37] programs to search EMBL and SwissProt database. X indicates a cycle in which no single amino acid could be identified with confidence. Peptide mass fingerprinting was carried out using MaxiSearch [28, 29] and MSMS searches were performed using Sequest [30].

<table>
<thead>
<tr>
<th>Spot number</th>
<th>MALDI-TOF-MS tryptic digest peaks</th>
<th>N-terminal sequence</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>875.5, 1031.7, 1070.6, 1086.6, 1099.5, 1107.3, 1118.7, 1191.8, 1216.8, 1537.8, 1735.2, 1891.4, 2356.6, 2378.7</td>
<td>METTVCPFFLM</td>
<td>HspH (B. japonicum) (64.2 % sequence coverage*)</td>
</tr>
<tr>
<td>2</td>
<td>936.4, 993.5, 1129.7, 1477.7, 1493.8, 1509.7, 2151.4, 2279.5, 2587.6, 2603.6</td>
<td>ALVTAQRA</td>
<td>Similar to 30S ribosomal protein S6 (Rhodobacter capsulatus)</td>
</tr>
<tr>
<td>3</td>
<td>1070.5, 1078.5, 1086.5, 1149.6, 1170.5, 1324.6, 1149.6, 1170.5, 1324.6, 1353.9, 1475.8, 1508.9, 1581.6, 1603.7, 1999.0, 3181.6, 3738.0</td>
<td></td>
<td>HspD (B. japonicum) (83.5 % sequence coverage*)</td>
</tr>
<tr>
<td>4</td>
<td>965.7, 982.8, 1176.0, 1266.1, 1651.2, 1849.7, 1942.6, 2212.9, 2525.3, 2809.5, 3341.2</td>
<td>MRTVDLPFL</td>
<td>Similar to HsP1</td>
</tr>
<tr>
<td>5</td>
<td>1036.6, 1109.4, 1265.7, 1323.8, 1383.8, 1607.7, 1648.7, 1811.1, 1911.2, 1995.2, 2264.8, 2086.8</td>
<td>AGTVQKLAMA</td>
<td>No match / homology</td>
</tr>
<tr>
<td>6</td>
<td>1070.7, 1086.7, 1108.7, 1177.8, 1193.8, 1268.9, 1291.8, 1307.8, 1324.9, 1346.9, 1476.0, 1573.1, 1633.0, 1995.5, 2355.6, 2724.0, 3040.3</td>
<td>MRTYDFAPL</td>
<td>HspP (B. japonicum) (65.4 % sequence coverage*)</td>
</tr>
<tr>
<td>7</td>
<td>862.7, 898.7, 931.6, 1149.8, 1170.7, 1258.8, 1395.9, 1303.8, 1524.9, 1562.9, 1676.1, 1491.9, 1592.0, 1602.9, 1986.4, 2538.9</td>
<td>MRTYDEPPWM</td>
<td>HspE (B. japonicum) (64.9 % sequence coverage*)</td>
</tr>
<tr>
<td>8</td>
<td>1179.6, 1191.8, 1277.8, 1307.8, 1328.8, 1476.0, 1523.7, 1709.1, 1839.4, 1853.3, 1870.2, 1995.4, 2212.7</td>
<td>MRTYLFTPLM</td>
<td>Similar to HspH</td>
</tr>
<tr>
<td>9</td>
<td>1035.5, 1082.5, 1103.6, 1199.8, 1340.0, 1383.8, 1441.4, 1476.0, 1898.7, 2228.9, 2921.1</td>
<td>MIKDLIPWN</td>
<td>HspC (B. japonicum) (47.6 % sequence coverage*)</td>
</tr>
<tr>
<td>10</td>
<td>1018.5, 1051.6, 1149.6, 1170.7, 1184.8, 1191.9, 1202.7, 1307.8, 1324.8, 1476.2, 1537.9, 1581.9, 1639.0, 1708.0, 1781.2, 1840.3, 2054.5, 2212.6</td>
<td>MRTYDFPSLM</td>
<td>Similar to HspD</td>
</tr>
<tr>
<td>11</td>
<td>806.3, 802.4, 898.3, 1045.5, 1060.5, 1074.5, 1096.5, 1112.4, 1155.5, 1178.5, 1194.6, 1208.5, 1224.5, 1233.5, 1246.5, 1599.7, 1621.6, 1637.7, 1778.0, 1802.0</td>
<td>NHFEPLHDVR</td>
<td>GroEL (B. japonicum) (65.3 % sequence coverage*)</td>
</tr>
<tr>
<td>12</td>
<td>825.4, 904.8, 990.6, 1035.8, 1207.0, 1263.8, 1481.8, 1666.9, 1868.1</td>
<td>VKREKDE</td>
<td>GroEL2 (B. japonicum) (22.1 % sequence coverage*)</td>
</tr>
<tr>
<td>13</td>
<td>592.5, 656.4, 722.5, 752.5, 826.8, 933.6, 1181.5, 1277.6, 1537.7, 1609.1, 1764.1, 1788.9, 1887.8, 2181.0, 2304.7</td>
<td>ADQLHSHODL</td>
<td>Similar to ribosomal protein L7/12 (bacteria subsp/Brucella melitensis)</td>
</tr>
<tr>
<td>14</td>
<td>1029.4, 1037.3, 1067.5, 1277.6, 1307.7, 1383.6, 1475.7, 1638.6, 1745.7, 1808.8, 1880.3, 1942.1, 1995.1, 2150.2, 2358.2, 2706.5, 2808.6, 2915.8, 3340.2, 3326.0, 3639.6</td>
<td>MNYFJW</td>
<td>No match / homology</td>
</tr>
<tr>
<td>15</td>
<td>466.3, 501.2, 514.4, 733.4, 860.1, 937.8, 1064.0, 1002.7, 1817.7, 12050.5, 3063.8, 3366.2</td>
<td>VXQI</td>
<td>No match / homology</td>
</tr>
<tr>
<td>16</td>
<td>532.2, 633.3, 659.4, 661.4, 700.5, 1049.5, 1133.8, 1226.7, 1323.6, 1308.7, 1435.0, 1510.9, 1518.9, 1816.9, 2245.3</td>
<td>XAKRPYKVs</td>
<td>double spot GroELa and b (Sequence coverage not calculated)</td>
</tr>
<tr>
<td>17</td>
<td>486.3, 604.3, 618.3, 633.3, 660.4, 683.3, 776.6, 854.5, 1019.6, 1081.4, 1232.6, 1293.4, 1383.2, 1561.6, 2302.1</td>
<td>XAKEYKVD</td>
<td>GroEL2 (22.3 % sequence coverage*)</td>
</tr>
<tr>
<td>18</td>
<td>475.3, 613.4, 633.5, 661.4, 775.5, 1049.6, 1123.4, 1222.7, 1393.6, 1533.7, 1561.6, 1640.5, 1928.9</td>
<td>XAKEVFSPVD</td>
<td>GroELa (65.3 % sequence coverage*)</td>
</tr>
<tr>
<td>19</td>
<td>502.5, 515.6, 700.4, 703.3, 777.3, 973.4, 1016.5, 1033.4, 1053.6, 1082.5, 1276.6, 1459.7, 1668.1, 2510.4, 2560.3, 2810.5, 2868.4, 4728.2</td>
<td>XNCVXDLGTV</td>
<td>DnaK (24.9 % sequence coverage)</td>
</tr>
</tbody>
</table>

*a Sequence coverage is the percentage of the sequence that is accounted for by the peptides generated by trypsin digestion.

*b Similarity is based on a combination of the sequence homology of the N-terminal sequence and the number of matching tryptic peptide masses.
on the HP G1000A Protein sequencer requires an initial yield after electrophoresis and blotting of around 2-3 pmol for successful sequencing, five times more than what was available on one gel. We therefore collected spots from multiple gels and focused the electrophoresis and blotting of around 2-3 pmol for successful digestion. The protein was then electroblotted onto a polyvinylidene difluoride membrane for either N-terminal sequencing or for digestion.

### Protein Identification: Peptide mass fingerprinting

Two thirds of the PVDF membrane was used for tryptic digestion. One µl of the protein digest was used for MALDI-TOF-MS analysis, and reproducible spectra were obtained at the 500 femtomole level for protein digests in the presence of β-octylglucopyranoside. The sets of peptide molecular masses obtained are given in Table 1. The data were used to search protein (SwissProt) and nucleic acid sequence (EMBL) databases using the program MassSearch [28, 29]. Figure 3 shows the MALDI-TOF mass spectrum of the trypsin digest of spot 11, which could be identified as GroESL. The top scoring proteins were analysed in a second pass, allowing for partial digestions and modifications such as oxidation, deamidation, carbamylation etc. The masses, which did not match, were used to search the database again to check if two or more proteins were present in the same spot. Eight spots could conclusively be identified as heat shock induced proteins, HspB, C, D, E and H and GroES, GroEL and DnaK (Table 1).

### Protein identification: Peptide fragment fingerprinting

After peptide mass fingerprinting, half of the remaining digest was subsequently used for HPLC-MS/MS peptide fragment fingerprinting. The samples were separated by reversed phase-HPLC and eluted by a gradient of acetonitrile in water into the triple stage quadrupole mass spectrometer for automated on-line HPLC MSMS data collection using a modified version of the Autofrag program as described in experimental procedures. Figure 4 shows the MS/MS spectrum of the tryptic peptide 1581.6 from spot 3. Database searching with the Sequest program [30] using the uninterpreted spectrum identified the protein conclusively as HspD. Spots 12 (GroESL), 16 (GroEL4 and GroEL4) and 17 (GroEL2) could be identified and the eight spots previously identified by peptide mass fingerprinting were confirmed (data not shown).

### Protein identification: Edman N-terminal sequencing

The remaining eight spots could not be conclusively identified by MS methods since the proteins were obviously not in the database. Therefore N-terminal Edman degradation was carried out with the remaining third of each protein blotted onto PVDF. Sequences were obtained from all spots (spot 16 showed a mixed sequence comprising GroEL4 and GroEL3). Three spots could be identified by sequence similarity as members of the Hsp family; spot 4 is homologous to HspD, spot 8 similar to HspL and spot 10 similar to HspD. In a database search, spot 2 showed a 60% identity in a 30 amino acid overlap to the 30S ribosomal protein S6 from *Rhodobacter capsulatus*. Spot 3 revealed a 75% identity in a 16 amino acid overlap to the 30S ribosomal protein S6 from *Rhodobacter capsulatus*. The remaining three spots, 5, 14, and 15 did not show any strong homology to known proteins and are therefore classified as novel.

### Kinetic analysis of the induction of heat shock protein expression

In order to obtain more information about the regulation of this large group of heat shock induced proteins, a kinetic analysis of the induction of eleven of the proteins was undertaken. The heat shock induced proteins that were identified fell into four classes: known sHsps, novel sHsp homologues, GroEL/GroES, DnaK proteins and unknowns. Of the ten sHsps, two, HspA and F, were not observed on these gels. The pI of theformer (8.4) is on the edge of the pH range of the first dimension used and the latter is probably too scarce since HspE, the product of the first gene of the *hspE* operon is barely visible. Three other spots, 5, 8, and 10 were too weak to obtain reproducible data as was HspE (spot 7), which was too close to the edge of the acidic range of the gels. The spots corresponding to the remaining four sHsps (HspB, C, D, H, 2 and 4), the GroESL proteins (GroEL2 and GroEL4 as well as GroES, and GroEL3) and DnaK were chosen for analysis. Eighty
Heat shock protein induction (Eur. J. Biochem. 263)

**Fig. 5. Kinetic analysis of the induction of some of the heat shock proteins.** The integrated optical density of some of the Coomassie Blue stained spots shown in Figure 2 are plotted against time. The values are given relative to the initial amount of protein present before heat shock. The mean error of the measurements was around ±20% for all points.

Two-dimensional gels were run and processed as described in experimental procedures. The integrated normalised OD values plotted in Figure 5 are the values relative to the initial spot density before heat shock. All the data are plotted on the same scale for ease of comparison. Clearly four different induction patterns can be seen. The largest induction is seen for GroES1, EL2 and EL4, which peak at 13 min before slowly decreasing to a steady state value (Figure 5D). A similar pattern is seen for HspD, H and 2 though the peak induction is clearly shifted to 14 min and the absolute increase is smaller (Figure 5B). HspB and 4 follow another pattern, increasing slowly to the steady state around 14 min without any overshoot. HspC is similar but shows a slight lag that is logical since it is the second protein to be made from the hspBC operon (Figure 5A). Finally, GroES1 and DnaK show a fourth type of induction, a slow steady rise to steady state, starting much earlier than the sHsp (Figure 5C).

**DISCUSSION**

The heat shock response of *B. japonicum* is a very complex one. This is in part due to the large number of proteins involved and in part to the variety of control mechanisms governing their expression (Table 2). Several members of heat shock proteins are known and have been characterised: Hsp70 (DnaK), the chaperonins (Hsp60) represented by at least five GroESL, Hsp40 (DnaJ) and the small heat shock proteins which consist of two separate classes. No members of the Hsp100 and Hsp90 heat shock proteins are so far known in *B. japonicum*. The *dnaKJ*, *groESL*, and *hspA-H* operons are under the control of the RNA polymerase transcription factors σ32 (the housekeeping σ32-like promoter binding factor), σ34, or at least two of the three versions of σ32 (RpoH1,2). RpoH2 is probably essential for growth under physiological conditions, supplying the cell with DnaK [31]. Under stress conditions a weak RpoH1-dependent expression of the *dnaK1* promoter was found [32]. The expression of the five *groESL* operons appears to be controlled by four mechanisms. Genetic studies revealed [16] a low constitutive expression of *groESL* and an oxygen-responsive NifA-dependent expression of *groESL*. The heat-induced expression of *groESL* and *groESL* is controlled by two separate mechanisms. RpoH1 is involved in the stress response, increasing the level of GroESL. The transcription of *groESL* and *groESL* is controlled by an inverted-repeat structure called CIRCF (controlling inverted repeat of chaperone expression, [33, 34]). A similar strategy to make a housekeeping promoter heat-inducible by superimposing a cis-acting regulatory DNA element is used in the case of the *hspA-H* operons. ROSE (repression of heat shock gene expression; [20]) serves as a repressor element during normal growth.

The identification of the proteins of Table 1 using the proteomic tools peptide mass fingerprinting and peptide fragment fingerprinting allowed the discrimination between GroES2 and GroES1. The two proteins possess similar pI and MW (GroES1: 7.97 / 11206.12 kDa and GroES2: 7.95 / 11177.08 kDa). Because of the almost identical pI and the only
Table 2. Induction characteristics of known heat shock proteins

<table>
<thead>
<tr>
<th>Gene</th>
<th>Promoter</th>
<th>mRNA level before HS</th>
<th>mRNA level after HS</th>
<th>Protein level before HS (OD)</th>
<th>Protein level after HS (OD)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>dnaK</td>
<td>$\sigma^{32}$</td>
<td>Low</td>
<td>x30</td>
<td>0.48</td>
<td>0.75</td>
<td>[31]</td>
</tr>
<tr>
<td>groEL2</td>
<td>$\sigma^{32}$</td>
<td>Zero</td>
<td>x150</td>
<td>El$_{1}$ not obs.</td>
<td>El$_{1}$ not obs.</td>
<td>[34]</td>
</tr>
<tr>
<td>groEL3</td>
<td>PH, CIRCE</td>
<td>Very low</td>
<td>x1</td>
<td>El$_{2}$ 0.52</td>
<td>El$_{2}$ &gt;3.5</td>
<td>[16]</td>
</tr>
<tr>
<td>groEL4</td>
<td>UAS, $\sigma^{41}$</td>
<td>Zero*</td>
<td>Zero*</td>
<td>El$_{2}$ 0.52</td>
<td>El$_{2}$ &gt;3.5</td>
<td>[16]</td>
</tr>
<tr>
<td>groFL2</td>
<td>PH, CIRCE</td>
<td>Low</td>
<td>x40</td>
<td>El$_{4}$ 0.62</td>
<td>El$_{4}$ 3.0</td>
<td>[34]</td>
</tr>
<tr>
<td>groEL5</td>
<td>PH, CIRCE</td>
<td>Very low</td>
<td>x10</td>
<td>El$_{4}$ not obs.</td>
<td>El$_{5}$ not obs.</td>
<td>[34]</td>
</tr>
<tr>
<td>hspD</td>
<td>PH, CIRCE</td>
<td>Very low</td>
<td>High$^b$</td>
<td>not obs.</td>
<td>0.41</td>
<td>[35]</td>
</tr>
<tr>
<td>hspE</td>
<td>PH, CIRCE</td>
<td>Very low</td>
<td>High$^b$</td>
<td>not obs.</td>
<td>0.32</td>
<td>[35]</td>
</tr>
<tr>
<td>hspF</td>
<td>PH, ROSE</td>
<td>Very low</td>
<td>High$^b$</td>
<td>not obs.</td>
<td>0.93</td>
<td>[20]</td>
</tr>
<tr>
<td>hspH</td>
<td>PH, ROSE</td>
<td>Very low</td>
<td>High$^b$</td>
<td>not obs.</td>
<td>0.35</td>
<td>[20]</td>
</tr>
</tbody>
</table>

* groEL3 genes are activated by the nitrogen fixation regulatory protein NifA at low oxygen conditions. No induction is observed under heat shock conditions. ^2 hard to quantify due to the very low basal level. ^3 protein levels in OD units after a 30 minutes heat shock at 43°C. The references refer to data concerning the mRNA levels. CIRCE, controlling inverted repeat of chaperone expression; PH, $\sigma^{34}$ dependent housekeeping promoter; a $\sigma^{34}$-like promoter; ROSE, repression of heat shock gene expression; UAS, upstream activator sequence, binding site for NifA; not obs., not observed;

The kinetics of DnaK protein synthesis shown in Figure 5C are in agreement with the results obtained previously by Minder et al. [31]. In this study it was demonstrated, that the heat-induced synthesis monitored by Western Blot and mRNA analysis shows a strong increase on transcriptional and translational level, which remains for at least 5 hours. The kinetic data differs to CIRCE-dependant transcription of B. japonicum groEL. Our studies are in perfect agreement with the results previously obtained for DnaK, showing a steady rise to a steady state. The induction kinetics for GroES$_{2}$, GroEL$_{2}$, GroEL$_{4}$ (CIRCE-dependant; Figure 5D) and the two sHsp classes, HspH, HspD, Hsp2 (ROSE-dependant; Figure 5B) and HspB, HspC, Hsp4 (ROSE-dependant; Figure 5A) show a clear different course. The kinetic of DnaK resembles instead the behaviour of GroES$_{2}$ induction that is also under $\sigma^{32}$ control.

The kinetic results obtained for the GroEL family fall into line with the regulatory control mechanism. GroES$_{2}$, GroEL$_{2}$ and GroEL$_{4}$ (as well as GroEL$_{4}$, though this is hard to prove since it appears as a double spot together with a second GroEL$_{2}$) show the same form of induction curve. The amount of protein increases rapidly and reaches a peak at around twelve minutes before decaying down to a fairly high steady-state level (Figure 5D). The respective groEL$_{2}$ operons have housekeeping $\sigma^{34}$-promoters upstream of a CIRCE element. As expected, all the genes are constitutively expressed at a low level due to the repression of the HrcA repressor binding to the CIRCE region and are strongly induced by heat shock (Minder, A. unpublished results). This has been previously shown at the mRNA level for groEL$_{2}$ and sHsps by primer extension studies, though the level of the GroEL$_{2}$ transcript was reported to be unchanged by heat shock. At the protein level this is clearly not the case, GroEL$_{2}$ and GroES$_{2}$ are expressed at fairly high levels at 28°C and after heat shock, there is a clear increase in expression. This could be due to an undescribed post-translational control mechanism or it could simply reflect that the GroEL$_{2}$ mRNA might be very unstable under stress conditions. The induction curve of GroES$_{2}$ is completely different to the other family members, being absent before heat shock and then slowly increasing to a steady-state after twenty minutes (Figure 5C). GroEL$_{2}$ is under the control of the heat shock transcription factor $\sigma^{32}$[34]; mRNA measurements indicate that groEL$_{2}$ is only present after heat shock but then at high levels. The protein results agree with the de novo synthesis, but do not show a massive increase in amount. Strangely GroEL$_{2}$ was not observed on these gels despite having a calculated pl and MW of 5.43 and 57.55 kDa respectively. It is in perfect agreement with previous results obtained with groEL$_{2}$-lacZ fusions and primer extension studies, that groEL$_{2}$ is sooner induced than groEL$_{3}$ [34]. The GroEL$_{2}$ operon is as expected not observed, because it is only induced under anaerobic conditions.

All of the sHsps operons have the same $\sigma^{34}$-type promoter and negatively cis-acting DNA element, termed ROSE. One would therefore expect all to show the same induction kinetics. In previous primer extension experiments [20, 35] a clear heat induction was found. As one can clearly see from Figure 5A and B, the curves obtained for the protein levels fall into two distinct types. The first (Figure 5A) shows a slow increase from zero to the steady state, whilst the second (Figure 5B) shows a distinct peak at fourteen minutes which drops rapidly thereafter to the
steady-state level. Both are distinct from the two GroEL induction curves. HspB and C are products of adjacent genes in the same operon and hence it is logical that they have the same kinetic response and that the curve of HspC which follows HspB shows a slight lag. Spot 4 represents a protein, which is homologous to HspE. The HspB operon shows a similar structure to the HspE operon, - two sHsps (an A and a B class Hsp) followed by a third open reading frame - so it may be logical that the HspE like protein shows a similar induction curve to HspB. The kinetics of HspE itself could not be determined because the protein migrated too close to the edge of the acidic range of the two-dimensional gel and was hence too variable between the gels. The HspD and H encoding genes both occur as isolated genes not in an operon and both show the same type of induction curve (Figure 5B) which is distinct from the hspBC operon. The kinetics of spot 2 induction follows the same pattern as HspD and H although its N-terminal sequence and the peptide fingerprint are completely different to typical sHsps. This supports the identification of spot 2 as a heat shock protein. The similarity of the N-terminal amino acids to the potential 3OS ribosomal protein S6 from Rhodobacter capsulatus makes it an interesting candidate for further studies presuming ribosomes as sensors of heat shock in B. japonicum. The identification of spot 13 as being similar to the 5OS ribosomal protein L7/L12 from Bracella abortus/Bracella melitensis underlines this hypothesis. It is interesting to note that this protein seems to be the binding site for several of the factors involved in protein synthesis and appears to be essential for accurate translation. A similar regulatory mechanism is suggested in E. coli, where ribosomes might be the primary sensor of conditions that evoke the heat shock response [36].

The use of reproducible and highly parallel processing of two-dimensional gels in combination with proteomic tools allows monitoring complex regulatory mechanisms by monitoring sets of proteins at one time. The proteome analysis proves to be a useful tool to group known and novel proteins into regulatory categories. This revealed a surprising interdependence between the small heat shock proteins and a ribosomal subunit, which makes it an interesting candidate for further studies. The analysis of gene function by following global changes in protein expression has proven to be a useful method to assign genes into functional groups. We show here that by not just using the start and end point of a process, but by analysing the time course of protein induction, functionally relevant information can be obtained which help to further classify genes into functionally related classes.

ACKNOWLEDGEMENTS

We are grateful to Huske Hennecke and Hans-Martin Fischer for generous support, continuous interest in our work, and helpful comments on the manuscript. This study was supported by grants from the Swiss National Foundation for Scientific Research and the Swiss Federal Institute of Technology, Zürich.

REFERENCES


Functional Proteomics

Heat shock protein induction (Eur. J. Biochem. 263)


of a functional nitrogenase in <em>Bradyrhizobium japonicum</em>, <em>Arch. Microbiol.</em> 171, 279-289
Curriculum vitae

Full name: Martin Christian Münchbach

Date of birth: August 29, 1970
Place of birth: Ludwigshafen/FRG
Nationality: German

Education:

1996-1999 Doctoral Thesis at ETH Zürich
Prof. Dr. Hauke Hennecke, Institute of Microbiology
Dr. Peter James, Protein Chemistry Laboratory
Development of Tools for Proteome Analysis

1994-1996 Undergraduate Studies in Biochemistry at ETH Zürich
Characterisation of the Liver Cytoskeleton and its Interaction with Calpain
Medal of ETH 1997 and Willi Studer Prize 1997

1991-1994 Studies in Biochemistry at University of Tübingen

1990-1991 Military Service, Würzburg

1981-1990 Gymnasium am Kaiserdom, Speyer
List of Publications


5. Urech, C., Kobv, S., Oppenheim, A., Münehbaeh, M., Hennecke, H. and Narberhaus, F. α52-type transcription factors that are resistant to degradation by the FtsH protease in vivo. Submitted to J Biol Chem


7. Santella, L., Kyoizuka, K., Hoving, S., Münehbaeh, M., Quadroni, M., James, P. and Carafoli, E. Breakdown of cytoskeletal proteins and calpain action during meiosis resumption by starfish oocytes. Submitted to Development


Acknowledgement

Writing the acknowledgements section is an especially difficult task, since it is hard to account for all the things during my dissertation. The following list will be for sure incomplete and will be more an attempt to thank all the people who helped to complete this thesis. Special thanks go to:

Prof. Hauke Hennecke, for accepting to be my supervisor and examiner, for continuous interest in my work and giving me the freedom becoming a protein chemist.

Prof. Ernesto Carafoli, for accepting to be my co-examiner and his recommendation to do my PhD in Proteomics.

Peter James: for his support and continuous help during the whole thesis, for continuous stimulation to progress, the sharing of his view of science and showing me how to survive the most chaotic situations happening in a laboratory. For his introduction to the real hot stuff and to the real black British humour.

Franz Narberhaus: for his support, continuous stimulation to progress, interest in my work and for introducing a protein chemist to the gene stuff.

Paola Dainese: for everything she told me about the making of two-dimensional gels and for providing the nice working atmosphere.

Werner Meyer: for his introduction to mass spectrometry and all the discussions while our coffee breaks.

Manfredo Quadroni: for his help in the last half of my thesis and helping me to survive the odd times during this stage.

To the rest of the A-floor: Chanty for providing the nice working atmosphere, Sjouke to show me how to play apeiron and Suzi for her help and support.

The K/J-floor, what has/is left: Holger, Fabrizio, Silvio, Timo, Li, Alessia, Joachim, Tony, Rolf, Linda and Paolo for his continuous support.

To all my friends, especially to KH, MAB, GL, CM, DM, AVS, DB, EM and to my family for their support and understanding.