Department of Physics, Chemistry and Biology

Master's Thesis

A metaproteomics-based method for environmental assessment: a pilot study

Henric Fröberg

2013-09-23

LITH-IFM-A-EX—13/2747—SE

Linköping University
INSTITUTE OF TECHNOLOGY

Linköping University Department of Physics, Chemistry and Biology
581 83 Linköping

Department of Physics, Chemistry and Biology

A metaproteomics-based method for environmental assessment: a pilot study

Henric Fröberg

Thesis work performed at IKE,
Faculty of Health Sciences, Linköping University

2013-09-23

Supervisor
Susana Cristobal

Examiner
Karin Enander

Linköping University
INSTITUTE OF TECHNOLOGY

Avdelning, institution
Division, Department

Chemistry
Department of Physics, Chemistry and Biology
Linköping University

Datum
Date

2013-09-23

Språk
Language

□ Svenska/Swedish
☒ Engelska/English

□ _______________

Rapporttyp
Report category

□ Licentiatavhandling
☒ Examensarbete
□ C-uppsats
□ D-uppsats
□ Övrig rapport

□ ____________

URL för elektronisk version

ISBN

ISRN: LITH-IFM-A-EX--13/2747--SE

Serietitel och serienummer
Title of series, numbering

ISSN
____

Titel
Title

A metaproteomics-based method for environmental assessment: a pilot study

Författare
Author

Henric Fröberg

Sammanfattning
Abstract

Metaproteomics, as a proteomic approach to analyse environmental samples, is a new and expanding field of research. The field promises new ways of determining the status of the organisms present in a sample, and could provide additional information compared to metagenomics. Being a novel field of research, robust methods and protocols have not yet been established. In this thesis, we examine several methods for a reliable extraction of protein from soil and periphyton samples. The extraction should preferably be fast, compatible with downstream analysis by mass spectrometry and extract proteins in proportion to their presence in the original sample.

A variety of methods and buffers were used to extract proteins from soil and periphyton samples. Concentration determinations showed that all of these methods extracted enough protein for further analysis. For purification and digestion of the samples, several methods were used. The purified samples were analysed on three different mass spectrometers, with the Orbitrap Velos Pro delivering the best results. The results were matched against four genomic and metagenomic databases for identification of proteins, of which the UniProt/SwissProt database gave the best result.

A maximum of 52 proteins were identified from periphyton samples when searching against UniProt/SwissProt with strict settings, of which the majority were highly conserved proteins. The main limitation for this type of work is currently the lack of proper metagenomic databases.

Nyckelord
Keyword
Environmental samples, metaproteomics, liquid chromatography, polyacrylamide gel electrophoresis, mass spectrometry

Abstract

Metaproteomics, as a proteomic approach to analyse environmental samples, is a new and expanding field of research. The field promises new ways of determining the status of the organisms present in a sample, and could provide additional information compared to metagenomics. Being a novel field of research, robust methods and protocols have not yet been established. In this thesis, we examine several methods for a reliable extraction of protein from soil and periphyton samples. The extraction should preferably be fast, compatible with downstream analysis by mass spectrometry and extract proteins in proportion to their presence in the original sample.

A variety of methods and buffers were used to extract proteins from soil and periphyton samples. Concentration determinations showed that all of these methods extracted enough protein for further analysis. For purification and digestion of the samples, several methods were used. The purified samples were analysed on three different mass spectrometers, with the Orbitrap Velos Pro delivering the best results. The results were matched against four genomic and metagenomic databases for identification of proteins, of which the UniProt/SwissProt database gave the best result.

A maximum of 52 proteins were identified from periphyton samples when searching against UniProt/SwissProt with strict settings, of which the majority were highly conserved proteins. The main limitation for this type of work is currently the lack of proper metagenomic databases.

Keywords
Environmental samples, metaproteomics, liquid chromatography, polyacrylamide gel electrophoresis, mass spectrometry.

Table of Contents

1 List of commonly used abbreviations

2D-LC	Two-dimensional LC
AA	Acrylamide
ACN	Acetonitrile
APS	Ammonium persulfate
BSA	Bovine serum albumin
CBB	Coomassie Brilliant Blue
CID	Collision-induced dissociation
COG	Cluster of orthogonal groups
DTT	Dithiothreitol
ESI	Electrospray ionization
FA	Formic acid
FASP	Filter aided sample preparation
FDR	False discovery rate
GC	Gas chromatography
HPLC	High pressure LC
IAA	Iodoacetic acid
IAM	Iodoacetamide
LC	Liquid chromatography
LC_x	Lethal concentration, the concentration of a chemical that kills x % of a population
LTQ	Linear Trap Quadrupole
MALDI	Matrix-assisted laser desorption/ionization
MS	Mass spectroscopy
MS/MS	Tandem MS
PAGE	Polyacrylamid gel electrophoresis
PMF	Peptide mass fingerprint
PTM	Post-translational modification
SDS	Sodium dodecyl sulfate
TCA	Trichloroacetic acid
TEMED	Tetramethylethylenediamine
TFA	Trifluoroacetic acid
TIC	Total ion current
TOF	Time of flight

2 Introduction

2.1 Background

There are several stressors affecting the environment today – both on a large scale and smaller, local scale. A factor such as global warming is expected to introduce a higher number of abiotic stressors such as extreme weather conditions. Environmental emergencies such as floodings might become more common, which can affect the biota [1]. Another stressor is pollution which, although becoming increasingly regulated by law, can cause a considerable impact on the environment. Heavy metals are expelled into the environment from mines and smelters, either by wastewater or discharge into the atmosphere [2]. Substances not seen as pollutants, such as food additives and drugs, are being expelled into the environment. Wastewater treatment plants were not built to handle the degradation of complex molecules such as pharmaceutical compounds. In addition, pharmaceutical compounds are often present in very low concentrations and have very diverse properties (size, solubility, hydrophobicity among others) making them difficult to remove. These compounds are likely harmless for humans, but it is still unknown how these substances affect the environment [3].

For many years, scientists have been able to determine what substances are present in an environmental sample. Using techniques from analytical chemistry such as mass spectrometry or liquid/gas chromatography, it has been possible to detect various substances present in a sample [4]. This has been an important part in classifying and deducing environmental effects. However, the combination of improved instrumentation, analytical methods and computational power opens opportunities to achieve more knowledge than before. For a proper determination of a substance's role, one would need to study its effect on the organisms present in the sample. This is possible for larger animals and has, in fact, been an important part in deducing the effects of pollutants, such as PCB and DDT in birds during the 1970s [5]. However, microbial organisms make up a larger biomass in total than other animals, and are also able to endure harsher conditions. As such, they are present almost everywhere. For a proper investigation of environmental effects, one would thus need to study the effects of pollutants on the microbiome as well [2].

A first study of this type on microbes was made possible by the introduction of metagenomics. Metagenomics emerged in the late 1990s and is the study of genes on a larger-than-organism sample, such as an environmental sample. With improved tools for sequencing such as Roche's 454, Illumina or SOLiD, metagenomics became an important tool for assessing the organisms present in a sample [6]. The area has enabled Craig Venter *et al* to discover what could be the fourth domain of life, by analysing small-subunit rRNA from samples taken during the Global Ocean Sampling Expedition [7], [8]. However, metagenomics can only provide an overview of the potential of the organisms – for example, the cellular pathways that can be activated. It does not provide any information regarding which pathways that actually are activated in the sample. Understanding the genetic potential of a sample is important, but an understanding of the function of the organisms is required for a more complete view of the effects of environmental stressors. The function of the organisms is assessed by studying the phenotype, or the expressed proteins, of the organisms. The expressed proteins give information about metabolizing schemes, which signal pathways are active, and are thus a measure of how the organisms are affected by the environmental stressors. The proteome will differ between two organisms of the same species but in different environments.

The complete set of proteins from an environmental sample is called the metaproteome and it is studied by metaproteomic technologies. The area has grown considerably during the last 20 years. Similar to metagenomics, this area of science has been made possible by the development in technology, mainly with respect to data management and mass spectrometry [9]. In the case of pollutants in the soil, metaproteomics would allow the study of the resident communities at a molecular level and the cellular changes of these organisms, rather than only the substances that are present. A study of how environmental and toxicological changes affect the organisms in the soil can provide new and more complex information for the assessment and early warning of environmental stressors.

However, as the field of metaproteomics is still in its infancy, standard procedures are not yet in place. Thus, all steps from sample processing, protein extraction, data integration and analysis have to be developed and integrated [10].

2.1.1 Proteomics

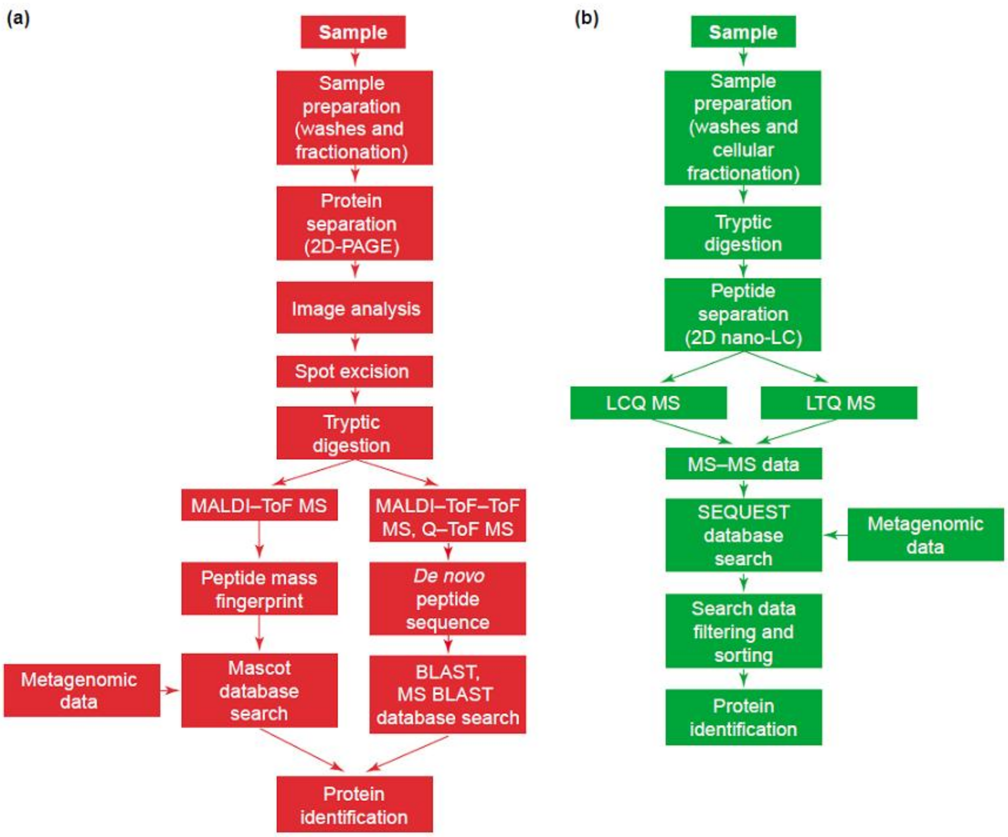
The word proteomics was used for the first time in the 1990s, and has come to mean the study of the complete set of proteins expressed in a cell, tissue or organism at a certain time. The area of proteomics has emerged thanks to advancements in several fields – protein separation, protein identification and information technology being the most important. The subject takes a global approach to proteins, studying complete cellular pathways and networks at the protein level [11], [12]. This differs from the previous, traditionalist approach, where it was common to study an isolated gene and its product. Proteomics has seen a massive increase since it was introduced, and is currently considered an important field for protein studies [12].

Protein separation in proteomics is usually carried out in one of two ways – by employing either gel electrophoresis or liquid chromatography. Separation of proteins is essential for subsequent analysis. In gel electrophoresis, mainly developed between 1960 and 1980, proteins are separated in one or two dimensions, usually depending on mass and intrinsic charge. The protein spots can then be excised from the gel and analysed. Liquid chromatography (LC) has appeared as another way to separate parts of a complex sample. In LC, a sample is introduced on a column, packed with a material with a certain property. The retention time of each compound depends on its affinity for the column's solid phase, which enables separation. Chromatographic separation has usually been carried out in one dimension but techniques have emerged for two-dimensional liquid chromatography as well [13]. As liquid chromatography has increased in popularity, the use of 2D-gels has decreased.

For protein identification, mass spectrometry is the method of choice for most scientists. Mass spectrometry for the analysis of peptides and amino acids was first used during the 1950's. It was at first a rudimentary technique, but it has since then been refined and developed. Mass spectrometry currently allows us to make accurate determination of peptide mass and sequence, even from a complex sample consisting of several hundred digested proteins. Mass spectrometry is commonly coupled to upstream LC for online sample handling with electrospray ionisation [11].

Improvements in information technology have enabled improved analysis and interpretation of the mass spectrometry data. In this case, that means comparing sequence information previously acquired with mass spectrometry data to determine which proteins were present in the sample. Some of the tools used for this are BLAST, FASTA and the newly developed Unipept [14]. These tools have allowed us to study the whole proteome of an organism.

The figure below shows two approaches for a proteomic workflow. Flowchart (a) depicts a workflow employing gel electrophoresis for protein separation, while flowchart (b) uses liquid chromatography. Both workflows use tryptic digestion, analysis on mass spectrometry instrumentation and database searches to identify the proteins present in a sample.


(a)
Sample
Sample preparation (washes and fractionation)
Protein separation (2D-PAGE)
Image analysis
Spot excision
Tryptic digestion
MALDI–ToF MS
MALDI–ToF–ToF MS, Q–ToF MS
Peptide mass fingerprint
De novo peptide sequence
Metagenomic data
Mascot database search
BLAST, MS BLAST database search
Protein identification
(b)
Sample
Sample preparation (washes and cellular fractionation)
Tryptic digestion
Peptide separation (2D nano-LC)
LCQ MS
LTQ MS
MS–MS data
SEQUEST database search
Metagenomic data
Search data filtering and sorting
Protein identification

- **Figure 1: Metaproteomic workflows, using either 2D-PAGE (a) or 2D-LC (b). Figure adapted from [15].**

Proteomics is currently used for several purposes. Some of these are protein cataloguing (identifying the proteins present in a sample, usually carried out with separation by LC or gels, followed by identification on MS), protein expression (comparing levels of protein expression between two samples) and PTM analysis [16]. It is also used for the study of protein-protein interaction and other purposes, which will not be used in this thesis.

The study of proteins has, however, turned out to be more difficult than the study of genes. There exist several reasons for this, of which the most prominent is the fact that proteins cannot be amplified (unlike genes, which can be amplified by PCR) [11]. Another reason is that knowledge of the genome is not enough to predict the complete set of proteins that will be expressed. An mRNA can be spliced, giving rise to variations of proteins, possibly with different biochemical functions. Proteins can also be modified after translation (a process which currently is difficult to predict), which can alter a protein's function, activity or stability. Thus, the expression of a single gene can result in several proteins [12].

2.1.2 Metaproteomics

Metaproteomics is, as defined by Wilmes and Bond:
"the large-scale characterization of the entire protein complement of environmental microbiota at a given point in time" [17].

A lot of knowledge has been gained from proteomic experiments, regarding protein functions, protein-protein interactions and disease biomarkers. However, until recently, the methods have mainly been applied to single cell cultures or tissue, because the samples have relatively low complexity. It is likely that other areas, such as microbial ecology, can benefit from the wider knowledge on cellular function that proteomics can provide [18].

Metaproteomics is an area of science that has grown over the past decade. It is an extension of proteomics. Metaproteomics, or environmental proteomics, is the study of the proteome on an environmental level – often a soil or water sample. The fact that the proteome does not come from a single organism has restricted the growth of this field – several factors have made metaproteomics studies difficult until recent years. For example, proteomic studies are often carried out with cell cultures. Extracting proteins from cell cultures is generally less complex than extracting proteins from a soil sample.

Microbial communities are present everywhere, also in habitats which are too harsh for higher-level species. These types of habitats include those with extremely low and high temperatures, high levels of radiation and low pH. Microbes have several important functions, and it is crucial to study these to better understand the microbes' roles in the environment. Some of these functions include converting carbon dioxide to organic molecules, biodegradation and fixation of nitrogen. Higher species rely on the microbes in many cases [19]. Earlier, metaproteomics researchers have tried to enrich the organisms present in the soil. This has, however, hampered the studies due to enrichment bias, or the fact that culturing techniques select for easily cultivable organisms [15]. It is estimated that between 90% and 99 % of the microorganisms in a soil sample are impossible to culture [18], [20]. Even when they can be isolated and grown in the laboratory, it is likely that they will not express the same characteristics as they did in their natural habitat [21].

Building on the data already provided by metagenomics, the next step is to elucidate functional change in the ecosystem upon exposure to stressors. This can be performed with metaproteomics techniques. The first metaproteomics studies were carried out on microbial communities from harsh conditions, since the few numbers of species present in the habitat led to a fairly low complexity [19]. However, metaproteomics studies are now carried out on a variety of samples.

2.1.2.1 Current metaproteomic research

Verberkmoes *et al* (2009) investigated the microbiome of the human gut, employing metaproteomic techniques. The aim of the investigation was to identify the proteins that could be confidently and reproducibly measured. Cells were extracted from human faecal samples. The cells were lysed and the proteins extracted. After desalting on RP C18 columns, the proteins were digested with trypsin, concentrated and filtered. The peptides were separated on 2D-LC (ammonium acetate salt pulses in one dimension and RP gradients in the other) before MS/MS analysis on a LTQ Orbitrap. The spectra were searched with SEQUEST against 4 databases (human metagenome, sequences of representatives of the gut microbiota and two decoy databases). Searching against the first database resulted in 600-900 non-redundant protein identifications (depending on the sample and run), while searching against the second database resulted in 970-1,340 identifications. The decoy databases were primarily used for controlling the ratio of false positives. The identified proteins were classified into clusters of orthogonal groups (COGs). Most detected proteins were involved in translation, carbohydrate metabolism or energy production. About 1/3 of the spectra belonged to human proteins. The relative protein abundance was estimated by calculating the normalized spectral abundance factor (NSAF). Human digestive proteins such as elastase, chymotrypsin C and salivary amylases were most common. According to Ram *et al* (2005) [22] proteins can be detected from populations representing at least 1 % of the community. This makes it likely that several populations and proteins are missed in a study like this [23].

The human intestinal microbiome has also been studied by Kolmeder *et al* (2012). Proteins extracted from faecal samples (taken from three subjects over six to twelve months) were separated on a 1D gel. The region expected to contain the most proteins (35-80 kDa) was cut out, the proteins digested and analysed on LC-MS/MS using an Orbitrap. The acquired spectra were searched against a total of five databases, ranging from metagenome databases to food databases. 1,790 proteins were identified with at least two peptides each. The core part, the microbial proteins that were identified in all three subjects at least once, consisted of 1,216 proteins. Functional analysis of these showed that metabolism of carbohydrates, nucleotides and amino acids were most common, reflecting the high metabolic activity of the microbiota. The team also discovered that while the presence of individual taxon could vary considerably, the overall composition of the proteome was more or less constant [24].

Rudney *et al* (2009) focused on the salivary microbiome. Peptides were separated three-dimensionally: first, the peptides were subjected to isoelectric focusing with a free-flow electrophoresis system. The most complex samples, as determined by MS/MS, were then subjected to two-dimensional separation, strong cation exchange (SCX) followed by reverse-phase (RP)-liquid chromatography coupled to MS/MS (LTQ). Microbial proteins were found by searching against SwissProt. A species was considered present if peptides matched to at least two proteins from that species, with at least one peptide unique for the species. Alternatively, if only a single protein from a species was identified, at least two unique peptides from the protein were needed for the species to be considered present. This led to the identification of 139 proteins from 34 different species. A COG analysis identified 4 major functional groups present in the sample: proteins were involved in translation, carbohydrate and amino acid transport and metabolism and energy production and conversion [25].

A study on Swedish twins tried to determine whether the metaproteome of the gut varied with disease. Six twin pairs were recruited to the study. The subjects were either healthy or had Crohn’s disease in either the small or large intestine. A superset of Swedish twins had already had their bacterial composition determined by 16S rRNA sampling. The current research focused on adding metagenomic and metaproteomic data. Cells were extracted from stool samples and proteins were extracted, digested and desalted. Peptides were separated two-dimensionally in a fashion similar to that of Verberkmoes et al [23].The spectra were matched to two databases using SEQUEST: one created from the metagenomic part of the study, where the microbial genome was sequenced, and one referred to as the human microbial isolate reference genome database (HMRG). Sequences for human proteins and common contaminants were added to both databases. Quantitative analysis of the proteins was performed using a label-free approach. The HMRG database gave the highest number of hits, between 1,930 and 2,900 hits for the three types of subjects (healthy and Crohn’s disease in small/large intestine). Considering that the first database was created by sequencing the metagenome of the subjects, this is a remarkable result. When comparing the COG categories of healthy subjects versus subjects with Crohn’s disease in the small intestine, several differences were found. Many categories related to energy production, transport and metabolism were significantly less represented in the diseased subjects compared to healthy ones [26].

Jagtap *et al* (2013) conducted research on the data treatment part of metaproteomics. Since proteomic experiments are commonly carried out on a single organism, it is easy to restrict the database searches to sequences from that particular organism. However, this approach is not possible in metaproteomics due to the often vast number of organisms in a sample. This necessitates the use of large databases, which has its drawbacks. Using a large database increases the risk for false positives (peptides that are identified but not present in the sample). Increasing the stringency (required to get high confidence results) however, increases the risk for false negatives (peptides that are not identified but present in the sample). Jagtap investigated a two-step database search to improve search results. In the first step, a search was carried out against a database. The proteins that are identified in this search (by at least one peptide) were used to create a new database. Searching against the new, smaller database resulted in more peptide-spectrum matches (PSM) of higher quality than before, thus reducing the amount of false negative hits. This was validated by spiking a sample, where the two-step database search method resulted in the confident identification of five times more peptides than the one-step search method [27].

Kan *et al* (2005) employed metaproteomic techniques to study the microbial diversity in an estuary in northeastern USA (Chesapeake Bay). Samples were collected from three spots in the bay, approximately 100 km from each other. The three proteomes were separated on 2D-gels. The gels of the middle and lower bay were fairly similar, while the gel with the upper bay proteome differed. A total 41 spots were excised from the gels and the proteins were digested. The peptides were analysed on MALDI-TOF, with 34 proteins giving spectra of high quality. A PMF search was performed using MASCOT, but no proteins were identified. With LC-MS/MS on a Q-TOF Ultima API-US followed by *de novo* sequencing and BLAST searches, the tentative identities of 3 proteins were found. It should be noted that this study was one of the first attempts at a metaproteomic approach to assessment of complex communities, and as such the instruments, data and software tools were not as good as they are today. It is likely that repeating this study with an Orbitrap, metagenomic databases and improved de novo sequencing tools would result in the identification of more proteins [28].

Another area which has received attention the last years is that of enhanced biological phosphorous removal, or EBPR. Microorganisms accumulate polyphosphate internally, thus removing phosphor from the wastewater. EBPR was studied by Wilmes *et al* (2008) [29]. Four sludges were taken from an EBPR reactor at different time points and concentration of phosphorus, three of them having good phosphor removal performance while the fourth sample performed poorly. The proteomes were separated on 2D-gel using IEF followed by SDS-PAGE. The gels from the three sludges that performed well were similar, while the fourth one differed. A total of 638 spots were common on the three gels from P-removing samples. Of these, 111 were excised, in-gel digested and analysed with MALDI-TOF-MS followed by a MASCOT search. PMF searches identified 38 proteins, while another 8 were identified with Q-TOF-MS/MS. Some identifications were redundant, leaving a total of 33 non-redundant protein identifications. The identified proteins were involved in PHA (polyhydroxyalkanoate) synthesis and fatty acid oxidation, glycogen degradation and synthesis, glyoxylate/TCA (tricarboxylic acid) cycles, phosphate transport and general stress response. Just as the research done by Kan *et al* (2005) [28], this study is a few years old and would most likely benefit from newer equipment and bioinformatic tools.

Wastewater treatment bioreactors have also been studied by Abram *et al* (2011). They studied the function of bioreactors at low temperatures. Since industrial wastewater is commonly discharged at low temperatures, it is common to heat it to a higher temperature before treatment. Bioreactors capable of working at a lower temperature would thus save the energy needed for heating the wastewater. The group extracted proteins from a bioreactor that had operated at 15 °C for 300 days. The proteins were separated on a 2D-gel using IEF and SDS-PAGE. A total of 388 spots were detected that could be reproduced. Of these, 70 were excised, digested and run on nanoLC-ESI-MS/MS on a Q-Star XL tandem mass spectrometer. The spectra were searched against NCBInr and TrEMBL, with at least 2 peptides required for the identification of a protein. A total of 18 non-redundant proteins were identified. 14 of these were involved in metabolism, mainly glycolysis and methanogenesis [30].

Bioreactors working at different temperatures were the focus of Siggins *et al* (2012) research. They studied how microbial diversity and protein expression in bioreactors varied with different temperatures and different amounts of TCE (trichloroethylene), a potentially carcinogenic compound used in industrial settings. Four bioreactors were operated for 235 days. They were either operated at 15 °C or 37 °C, as well as with or without TCE (60 mg/l). Various analyses were conducted alongside the metaproteomic analysis. The proteins, extracted from the biomass using sonication, were separated on a 2D-gel employing IEF combined with SDS-PAGE. Protein spots where the intensity varied more than two-fold between two different reactors were excised and analysed using nLC-ESI-MS/MS. Acquired spectra were searched against the NCBInr database using Mascot, with a minimum of two peptides required for protein identification. A total of 93 spots were excised, which led to the identification of 27 unique proteins. Identified proteins were involved in acetate and ethanol metabolism, as well as glyoxylate degradation. Half of the identified proteins belonged to species in the Proteobacteria phylum. Methyl malonyl-CoA mutase was upregulated 24-fold in the presence of TCE in the warm reactor, indicating that the methyl malonyl-pathway is active under these conditions [31].

Another environmental issue is the presence of vinyl chloride (VC), a known human carcinogen, at certain industrial sites. By investigating the microbial diversity and proteome of organisms capable of degrading VC, Chuang *et al* (2010) aimed at discovering protein biomarkers for these types of organisms. The microcosms[*] were created by incubating VC-contaminated groundwater with mineral salts and trace metals for 60 days. DNA and proteins were subsequently extracted from the samples. All etheneotrophic and VC-assimilating bacteria discovered so far employ the enzymes alkene monooxygenase and epoxyalkane: coenzyme M transferase. The genes for these (EtnC and EtnE) were thus amplified and sequenced. Proteins were separated either on 1D-SDS-PAGE or SCX-LC, followed by analysis on ESI-MS/MS. In all samples where either EtnC or EtnE genes or their corresponding proteins were found, etheneotrophic bacteria were present. This indicates that these genes or their proteins are appropriate biomarkers for these bacteria [32].

Metaproteomics on a grander scale has been applied by Morris *et al* (2010), in their investigation of the membrane proteins of microbes in coastal and open ocean waters in the south Atlantic. Membrane proteins were chosen because of their involvement in nutrient transport and energy transduction. 5 samples were taken in open waters and 5 samples were taken off the coast of southern Africa. After filtering, the cells were lysed and membrane proteins extracted. The proteins were digested with trypsin and analysed with LC-ESI-MS/MS on an LTQ-Orbitrap. A SEQUEST search against the genomic database from the Global Ocean Sampling (GOS) was performed. Most of the identified proteins were related to transport, had unknown functions, or were uncharacterized outer membrane proteins. Viral proteins were identified in all samples. The authors also performed a deeper study, consisting of 60 MS/MS runs on a single sample. Even with this many runs, only 238 of 3,639 proteins could be identified with more than one peptide. This demonstrates that metaproteomics is still a novel field. The authors identified 6.2 times more peptides when searching against the GOS metagenomic library than when searching against the GenBank nonredundant databases, demonstrating the importance of having a database that reflects the organisms that could exist in the sample [33].

2.1.3 Environmental assessment

The area of metaproteomics can open a new era in the environmental assessment providing insights both at the functional and the systemic level. This thorough assessment of environmental status has never been possible before. Traditional evaluation of biodiversity changes required a tremendous amount of manual work, inspecting species under a microscope. This is why although still in its infancy, the field of metaproteomics has high expectations.

Metaproteomics enables on the one hand the discovery and study of biomarkers in an environmental sample, and on the other hand the evaluation of changes in biodiversity. A biomarker is a compound which can be measured as an indication of an organism’s state. Biomarkers are often used to assess the effect of an organism on exposure to a substance, or to detect the difference between two organisms in different states (such as healthy/diseased or treated/non-treated) [34]. Earlier, with less advanced technology, the state of an environmental sample could be assessed with a battery of individual assays. These assessments commonly measured various physical and chemical properties, such as pH, concentration of metal ions, membrane permeability or variation in enzymatic activity. Some of these assessments are included in the standardized environmental protocols and they give a few hints on the status of the microbial communities in the sample [35], [36], but those assays are not always very robust against biotic and abiotic factors.

[*] Artificial ecosystems, used to study ecosystems under controlled conditions

2.1.4 Ecotoxicology

Ecotoxicology is the science of contaminants in the biosphere and their effects on constituents of the biosphere, including humans [37].

The area of ecotoxicology emerged during the middle part of the 20th century and is an extension of toxicology, or the study of adverse effects of chemicals on living organisms [38]. Ecotoxicology provides a wider perspective on chemical substances' adverse effects by studying how it affects organisms at levels of population and ecosystem. The area received attention by the publishment of Rachel Carson's book *Silent Spring* [38], drawing attention to the effect of accumulated pesticides on animal wildlife. The basis of ecotoxicology is to provide methods for assessing chemicals' effects on ecosystems and a foundation for how to manage them [39]. The origin in toxicology has not been without problems, however. Toxicological studies are commonly carried out in a lab, with well-defined protocols where a single species is exposed to a single/few toxicants. This is an accepted way of performing studies on e.g. a drug's adverse effects in animals and humans and is common in the initial steps of drug development. For assessment of environmental status, this way is not ideal. Environmental samples are often complex, where many species are exposed to a broad combination of physical, chemical and biological stressors. For evaluating a toxicant's effect on an ecosystem, it is necessary to extrapolate from the lab results to the real-life ecosystem [38], [39].

It is in this perspective that proteomics, and metaproteomics, can play an important role. Ecotoxicology is mainly based on *in vitro* experiments and extrapolation of the results from an isolated laboratory environment to complex environments. Proteomics enables us to study the samples *in vivo*, with no extrapolation of the results. The protein expression profile in an organism can be used to assess its response to environmental stress. Proteins which become up- or downregulated indicate which pathways are modified as the organism adapts to the change [40].

2.1.4.1 Current ecotoxicoproteomic research

Research in this area has included the study of molecular endpoints such as protein level, instead of lethality commonly used in ecotoxicology studies. Gündel *et al* (2012) studied the effects of phenanthrene on zebrafish embryos, with concentrations ranging from 1 % of LC_{50} to LC_{20}. The proteome was separated on 2D-gels using IEF for first-dimension separation and SDS-PAGE for the second dimension. A total of 713 spots were identified and 89 spots that were differentially regulated were excised. Using nanoLC-ESI-MS/MS on a LTQ Orbitrap XL, 21 proteins could be identified. They could thus create a protein expression profile for the response to phenanthrene. Some of the identified proteins were vitellogenin, where an up-regulation was previously connected to endocrine disruption, and structural proteins, where down-regulations has been seen as an indication for cytotoxicity [41].

A similar approach was taken by Dorts *et al* (2012). They studied how perfluorooctane sulfonate (PFOS) affects protein expression in the gill tissue of *Cottus gobio*, a candidate sentinel species. PFOS accumulates in the food chain and is associated with hepatotoxicity and reproductive toxicity in fish. Protein separation was performed on 2D-gels, followed by MS for identification of the proteins. Out of the 20 identified proteins, only 3 displayed common trends in expression in response to the various concentrations of PFOS. The remaining proteins were expressed differently at the different levels of PFOS. The identified proteins were involved in a variety of cellular functions, including the general stress response and energy metabolism [42].

This methodology has also been applied on soil samples. Wang *et al* (2010) studied the effects of cadmium exposure to earthworms, *Eisenia fetida*. The earthworms were exposed to an environment containing 80 mg $CdCl_2$ per kg soil for up to 28 days. Proteins were extracted and subsequently separated on 2D-gels. Spots corresponding to over- or underexpressed proteins were excised, digested and identified with MALDI-TOF/TOF-MS. Of 143 proteins being significantly over- or underexpressed at least once, 56 were identified. Of these, 28 were upregulated and 28 were downregulated. These proteins were involved in a variety of cell functions. 41 % of the regulated (both up- and downregulated) proteins were identified as related to metabolism. Other proteins were related to stress and defense response and translation. With Cd being a hazardous heavy metal, this research provides a way to understand how organisms cope with Cd exposure [40].

Cadmium exposure was also investigated by Choi and Ha (2009). They studied the effects of Cd exposure on globin mRNA, hemolymph protein expression and total Hb content in *Chironomus riparius*, a nonbiting midge. The study was performed with RT-PCR analysis of mRNA and 1D- and 2D-gels for protein expression analysis. *C. riparius* was subjected to 0.1 %, 1 % and 10 % of LC_{50}, approximately 210 mg/L. The proteins were extracted and analyzed on 1D-gels (using PAGE and IEF) as well as on 2D-gels. On the 2D-gels, the expression levels for 14 proteins differed. All of these proteins were globin proteins, and all but 2 were downregulated upon exposure to Cd [43].

The effects on the proteome from another pollutant, polychlorinated biphenyls (PCB:s), was investigated by Leroy *et al* (2010). They studied how the PCB:s CB77 and CB169 affected the freshwater invertebrate *Gammarus pulex*. *G. pulex* was exposed to aqueous solutions of the two compounds. The extracted proteins were separated on a 2D-gel and analysed by MALDI-TOF/TOF. Of the 560 protein spots visible, 21 exhibited large differences compared to the control group and 14 of these were identified. In general, proteins related to amino acid metabolic pathways were downregulated, while proteins related to the cytoskeleton were upregulated [44].

2.2 Aim of the project

The aim of the project is to develop a new method for extracting protein from different sources. In this work, soil and periphyton will be the protein sources. The soil comes from islands in Stockholm's archipelago and the periphyton comes from the river Zadorra (Álava province, Spain).

The short-term goal is to develop a robust method, with reproducible results, for extracting proteins from soil and periphyton samples and identify them using mass spectrometry. The method should preferably be as little discriminating as possible, meaning that the amount of each extracted protein should be proportional to the amount in the soil (i.e., no preference toward extracting e.g. hydrophilic proteins). After extraction, methods for purification, digestion and mass spectrometry analysis of the proteins will be researched. The last step will be identification of proteins by comparing mass spectrometry data to genomic and metagenomic databases.

The long-term goal is to apply the developed method on sets of samples treated in different ways, to determine protein expression profiles and finding suitable biomarkers. This is not a goal of the current thesis work, but the work presented here will hopefully be used to achieve this long-term goal.

The main delimitation of this project is time. Since it is a pilot study, there are no currently accepted methods for how to proceed. With the given time, there are a limited number of possible methods that can be tested. The lab has access to good instrumentation and skilled co-workers, leaving time as the major delimitation for what is possible to achieve with this work.

3 System and Process

During the first week, the project was planned according to the following chart:

Week	Task
1	**Start of thesis work**, literature study
2	Evaluation of protocols for the extraction, separation, purification and digestion of proteins
3	Evaluation of MS/MS protocols
4-5	Extract, separate, purify, digest and run proteins from all samples on MS/MS
6	**Introduction finished**
7-8	Analyse MS/MS data to determine which peptides are present
9	**Half time report**
10-11	Determine which proteins the peptides belong to
12	**Laboratory work finished**
13-14	Report writing
15	**Preliminary report sent to examiner**
16	**Report sent to opponent**
17	Report writing
18-19	Presentation + finish report
20	**Report sent to examiner**

The thesis work was initially planned to be carried out during the spring of 2013, beginning in early January and finishing in June. The aim of the literature study was twofold: first, to find content providing the theoretical background for topics I needed to know more about. Second, to find methods for protein extraction from published papers with similar aim. If appropriate and suitable methods were found, these were to be evaluated during the second week of work. The extracted peptides would then be run on the mass spectrometers the department has access to, to evaluate which one that provides the best result. Following this, the selected methods would be systematically applied to a set of samples treated in different ways. In parallel, report writing would start, beginning with the Introduction and Theory chapters. After all samples had been run on the mass spectrometer, analysis would then begin to determine which peptides that were present in the sample. Approximately halfway throughout the work, a half-time meeting was to be held with the examiner to determine whether the project was going in the right direction and going to be finished on time. The remaining laboratory work would be to determine which proteins the peptides belonged to, and to try to draw conclusions from this. The time remaining was to be spent writing the remaining parts of the report. The presentation was to be held in early June.

4 Theory

4.1 Protein Extraction

4.1.1 From soil samples

For protein extraction from soil samples, two methods will be used. Both methods use SDS as a detergent, which breaks interactions between proteins, lyses cell walls and prevents protein aggregation. For improved protein extraction, one method uses phenol, while the other boils the samples. Phenol was first used for purifying carbohydrates and nucleic acids and removing the unwanted proteins. During the past few years phenol extraction has begun to see its use in protein purification as well. A mixture of phenol and water is added to the sample. After extraction with vortexing and sonication the nucleic acids migrate to the water phase and the proteins to the phenol phase. An extraction with phenol is sometimes carried out with sucrose in the water. This creates a phase inversion as the water phase will be heavier than the phenol phase, facilitating extraction of the phenol phase [45].

4.1.2 From periphyton samples

A variety of methods will be used for extraction from periphyton samples. For cell lysis, two methods will be employed: grinding in liquid nitrogen or in a polytron. Both methods will force the cell membranes to rupture, and the aim is to compare the protein extraction capability. The buffers that will be used are composed of various detergents, surfactants and reducing agents.

4.2 Measurement of protein concentration

The Bradford assay was popularized in an article by Marion Bradford at the University of Georgia published in 1976. The assay has become a widely acknowledged method for determining protein concentration in a sample [46], [47].

The basis for the assay is the fact that certain dyes can bind to protein. In this assay, Coomassie Brilliant Blue G-250 is used. Upon binding, the absorption maximum changes from 365 to 595 nm. This can be seen as a colour change from red to blue. The protein-dye complex has a high extinction coefficient which leads to a good sensitivity in the measurement [48].

The method has some known drawbacks [46], however, they are not serious enough to motivate the use of another method. The work done in this thesis is mainly exploratory and it is not necessary to determine the protein concentration with high accuracy. In addition, it is likely that an estimated 50 % of proteins can be lost during the extraction, reducing the need for accurate concentration measurements [49].

4.3 Enzymatic digestion

Trypsine is a serine protease, cleaving proteins at the C-terminal side of Arg and Lys residues. It is common to use peptides for mass spectrometry analysis. Trypsin digestion is common before mass spectrometry since it only cleaves at two sites. This reduces computational needs and gives peptides which, most of the time, are of an appropriate length for MS with good ionization and fragmentation [50]. In addition, trypsin is generally quite stable and does not require as specific conditions as other proteases [51].

For the enzyme to work as efficiently as possible, the protein sample needs to be denatured, its disulfide bonds reduced, and the resulting cysteines alkylated. This is done to prevent them from forming disulfide bonds again. The denaturation is often carried out together with the reduction by the reagent DTT. The alkylation is often carried out with iodoacetamide (IAM) or iodoacetic acid (IAA) [50].

4.4 Liquid chromatography

Liquid chromatography is a technique used to separate compounds in a liquid phase. Liquid chromatography is commonly employed to identify the compounds in a sample, to quantify them or to purify the sample. In liquid chromatography, a mixture of molecules is fed onto the column. The chromatograph consists of two phases, a fixed stationary phase and a mobile, flowing phase. During elution, a gradient is often achieved by allowing the mobile phase to increase in hydrophobic or hydrophilic strength over time. Different compounds have different affinities for the stationary phase and will thus have varying retention times. There are different kinds of liquid chromatography, among them affinity chromatography, reverse-phase chromatography and ion exchange chromatography [16].

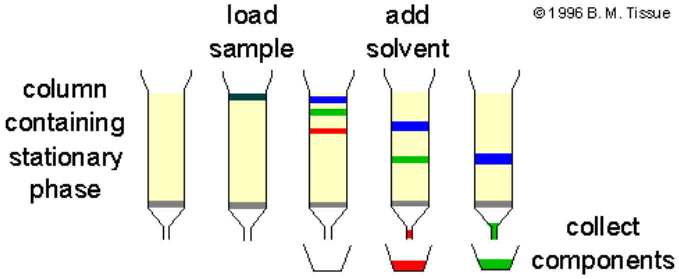


Figure 1 - A simple column for liquid chromatography. The sample, containing a mixture of substances, is applied. With different affinities for the stationary phase, the substances elute at different times. Figure adapted from [52].

4.4.1 Reverse-phase liquid chromatography – RP-LC

In reverse-phase chromatography, the stationary phase is non-polar. This is opposite to normal phase chromatography, where the stationary phase is polar. The stationary phase usually consists of carbohydrate chains of 4-18 carbon atoms immobilized on silica particles. To achieve a gradient elution, the amount of organic, non-polar compounds in the mobile phase is gradually increased. This separates the peptides according to their hydrophobicity. RP-HPLC is also quasi-mass dependent, since the hydrophobicity and thus retention on the column generally increases with mass. The method gives a high resolution and is often used in between tryptic digestion and mass spectrometry [16].

4.5 Gels

Gels are another way to separate proteins or genetic material for analysis. The advantage of gels is that the result can be inspected visually.

SDS-PAGE (SDS-polyacrylamide gel electrophoresis) is by far the most commonly used gel technique for separation of proteins. SDS-PAGE separates proteins in one dimension. By combining it with IEF (isoelectric focusing), proteins can be separated in two dimensions which is useful when working with complex protein samples. Regardless of the actual type used, gels build upon the phenomenon of electrophoresis, or the fact that a charged particle will move if an electric field is applied.

SDS, or sodium dodecyl sulfate, is a detergent used for preparing proteins for the gel. It denatures the protein and binds stoichometrically to the peptide backbone. SDS gives the molecule a negative charge which is, for all purposes, proportional to its mass. Any intrinsic charges the protein carries are dwarfed. The sample is then loaded on a gel, consisting of a mixture of acrylamide and *bis*-acrylamide. *Bis*-acrylamide works as a link between the acrylamide molecules. The composition of the gel determines the size of the pores. The higher the concentration of acrylamide, the smaller the pore size (to a certain extent). The pore size, in turn, determines how quickly the proteins will pass through it. A gel with small pore size will only allow the smallest proteins to pass through quickly, while the larger proteins will migrate very slowly. On the other hand, a gel with a large pore size will sieve the larger proteins well, while smaller proteins might pass through too quickly [16]. It is common to use gels with a percentage of acrylamide between 6 and 15, depending on the size of the proteins of interest.

A positive voltage is applied at the opposite end of where the proteins are loaded. The negatively charged proteins will migrate toward the positive anode, with a speed depending on their charge (and thus their size) and the pore size of the gel. It is common to include a reference sample, with proteins of known molecular masses, to the gel. The reference sample, or ladder, can then be used to approximate the masses of the unknown proteins [16].

4.5.1 Gel staining

For detection of the proteins in the gel, the gel can be stained. There are various stains, each with their own advantages. Most stains bind to the proteins and not the gel (so called positive staining) but there exists a few that stain the surrounding gel and not the proteins (negative staining). Stains can be introduced into the gel before or after electrophoresis. The stains can be used to detect all proteins in a sample or only specific proteins. Specific PTM:s can also be detected with some stains. Stains can be radioactive, fluorescent or visible in normal light [16], [53].

In this thesis, Coomassie brilliant blue (CBB) and silver staining was used. CBB belongs to a group of positive stains that can be detected in visible light. CBB was introduced in 1963 and was originally used to dye textiles. Under acidic conditions, CBB binds to the amino groups of the proteins. Staining with CBB is commonly done in a solution of a weak acid, alcohol and water. Depending on the protocol, the staining process can take from a few minutes to overnight. The CBB that does not bind to proteins can be washed out of the gel, a process known as destaining. This leaves a gel with stained protein bands. Destaining is performed with the same solution as for staining, except for the CBB dye. The time for destaining varies between minutes and overnight. The stained bands can be excised from the gel and destained for later analysis, since CBB does not interfere with downstream digestion and mass spectrometry [53].

Silver staining is more sensitive than Coomassie staining by at least an order of magnitude. Silver staining was previously incompatible with downstream mass spectrometry, but protocols have been developed for MS-compatible silver staining. A disadvantage with silver staining is that it has a narrow dynamic range. This is a problem when performing quantitative analysis, but that is of little importance to this work [16].

4.6 Mass spectrometry

The idea behind mass spectrometry is that ions can be separated according to their m/z ratio – that is, their mass over charge ratio. The m/z ratios of the ions can then be used to identify the molecules present in the original sample. Further development has even led to the ability to sequence proteins and peptides using an extension of mass spectrometry known as tandem MS or MS/MS [16].

A mass spectrometer consists of three parts: ion generation, ion separation and ion detection. These will be explained below, together with some technical aspects of mass spectrometers such as accuracy and resolution.

4.6.1 Ion generation

Of the methods currently used to ionize biological samples such as peptides and proteins, MALDI and ESI are by far the most common. Several variants of these methods have developed, but the general idea is the same [54].

Matrix-assisted laser desorption-ionization (MALDI) was developed during the 1980's. MALDI is performed in two steps. In the first step, the sample is dissolved in the so-called matrix. The matrix is a substance which can absorb the energy from the laser pulse and transfer some of it to the analytes. A drop of matrix-dissolved sample is then allowed to dry on a metal plate. In the second step, the plate is inserted into the MS instrument. The crystals are fired upon with short pulses of laser. This sublimates the matrix and ionizes the analytes. The analytes can then be separated and detected. MALDI is commonly used with a time-of-flight (TOF) ion separator [54].

Electrospray Ionisation (ESI) is the other way to produce ions for mass spectrometry. ESI has a high sensitivity and has the ability to ionize large samples such as peptides. In addition, it is easy to couple to upstreams separation techniques such as HPLC, allowing for high-throughput analysis. The principle behind ESI is that ion-containing droplets will decrease in size if they are situated in an electrical field. If the molecule one wants to analyse is present in the droplet (e.g. a peptide), the droplet will decrease in size, eventually leaving only the ionised molecule [54], [55].

This is achieved in the following way: the solution with the analyte/analytes is transported in a metallic capillary with low flux. By applying a potential difference of 3-6 kV between the capillary and a counter-electrode 3-20 mm away, an electric field is obtained. Upon leaving the capillary, the droplets will be charged and transported towards the counter-electrode in the shape of a lens. The droplets are often passed through some kind of heating source, usually a heated capillary or an inert gas such as nitrogen, where solvent molecules are removed. The isolated analyte can then be transported toward the mass spectrometer. See figure 2 for an illustration of ESI.

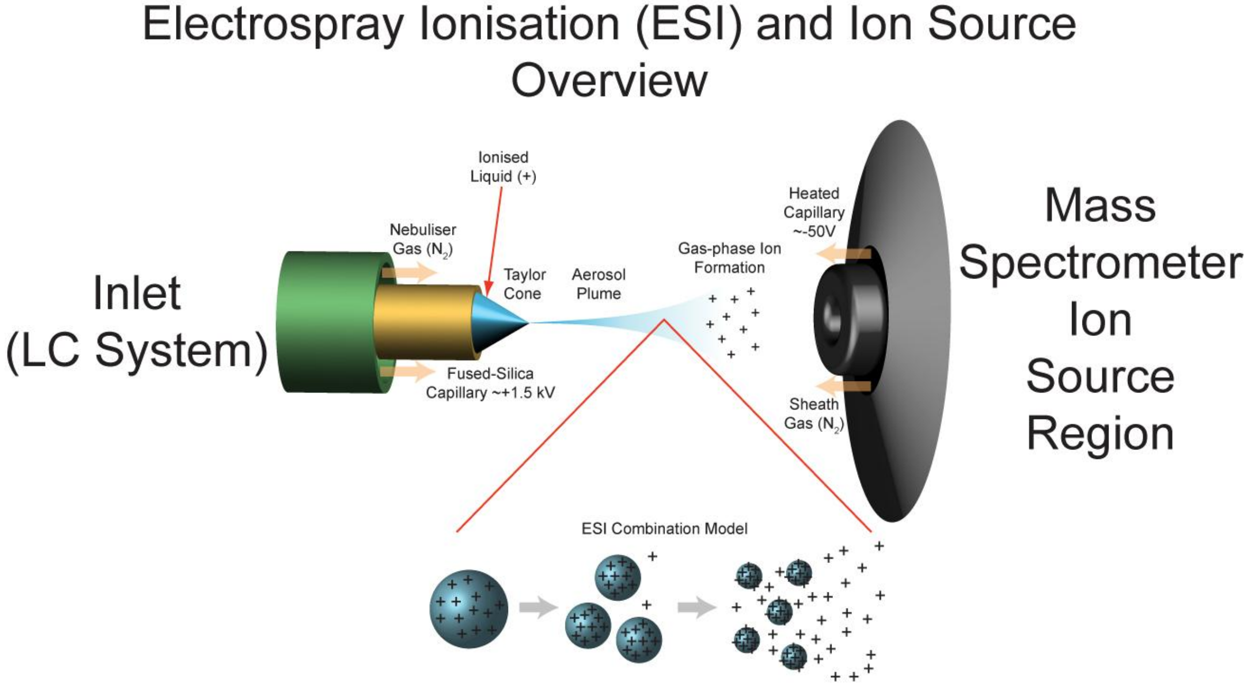


Figure 2 - Electrospray Ionisation. Figure adapted from [56].

There are two models that explain why the droplet decreases in size. The first, the charge residue model (CRM), states that as the solvent evaporates from the droplet, the droplet will become smaller but retain the same charge since no ions have evaporated. When the droplet reaches a certain size known as the Rayleigh limit, the repulsion between the charges will be too large, effectively splitting the droplet into smaller droplets. This will continue until all that is left is the analyte. The second model, the ion evaporation model (IEM), states that when the droplet becomes small enough, ions on the droplet's surface will be pushed out into the gas phase, decreasing the droplet's charge [57].

There is a disagreement about which of these models that most correctly describes the origin of the ions. According to Banerjee and Mazumdar, it seems that the IEM can explain the creation of the gas phase ions when the ion is small, while the CRM can explain the creation of larger ions [55].

4.6.2 Ion Separation

After generation, the ions need to be separated before they can be detected. These ion separators, also known as mass analyzers, can commonly be divided into two classes:

- Scanning analyzers, only allowing ions of a specific m/z ratio to pass through at any given time. A quadrupole is an example of a scanning analyzer.
- Simultaneous transmission analyzers: all ions are allowed to pass through the analyzer at the same time. The method of detection varies between the instruments. Examples of mass analyzers of this type are TOF instruments, the ion cyclotron resonance instrument and the Orbitrap [54].

The mass analyzers deploy different ways to separate the ions (e.g. kinetic energy, velocity, rotational frequency) but in the end, these all depend upon the m/z ratio of the ions [54].

The quadrupole is made of four usually circular metallic rods. There also exists hexapoles and octapoles, with six and eight poles, respectively. To two opposite rods, the same potential difference will be applied. The potential difference is a combination of direct current (commonly 500 – 2000 V) and alternating current. The positively charged rods will act as a low pass filter, only allowing ions with an m/z value higher than a certain limit to pass through. Similarly, the negatively charged rods will work as a high pass filter. Together, these filters create a window that only lets through ions of a certain m/z ratio. By varying either the direct or the alternating current, the window moves. By combining this with a proper detector, one can see at what m/z values ions passed through (and thus were present in the sample) [54].

For an ion, there are some combinations of direct and alternating current which allows it to be stable and pass through the quadrupole. By increasing the voltages from 0 to some arbitrary upper limit a scan can be performed. There are two ways of performing this scan: constant resolution scan or unit mass scan, with the latter being more common. In unit mass scan (or constant peak width scan), a higher resolution is achieved, at the cost of less ions reaching the detector [58].

An ion trap is either two-dimensional or three-dimensional. A 2D ion trap is similar to the quadrupole, but in addition to the four rods it has two additional lenses at the end of the rods. These lenses are repelling the ions of a certain charge, resulting in ions that are enclosed in the trap. A 3D ion trap is a quadrupole bent around itself. Two-dimensional traps have a higher trapping capacity than three-dimensional traps. After trapping the ions, the ions are expelled from the trap either axially (through one of the lenses) or radially (through one or more of the rods). An ion trap can be used either as an analyzer, or as a trap by creating potential wells along the electrodes' axis, storing the ions inside [54].

In a TOF instrument, the generated ions are accelerated from the plate due to a potential difference. The potential energy is thus converted to kinetic energy. The separation of ions is then based on the ions' velocities. An ion with a lower m/z ratio will reach the detector faster than an ion with a higher m/z. A TOF instrument requires that the ions are generated during a short time span. Because of this, MALDI is preferred while e.g. ESI is impossible to use. TOF instruments generally have a very high sensitivity [54].

4.6.2.1 Orbitrap

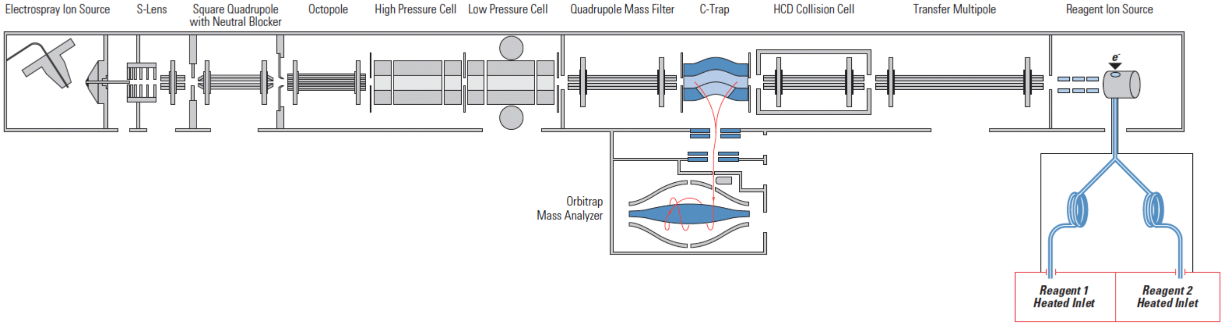


Figure 2 - A schematic picture of the Orbitrap Velos Pro

The Orbitrap was invented by Alexander Makarov during the 1990s-2000s. The first commercial instrument was introduced on the market in 2005. It has since then become one of the most widely used instruments due to its high resolution and mass accuracy. The Orbitrap is shaped as seen above: a spindle-like centre electrode surrounded by an outer electrode [59], [60].

The Orbitrap and the ion cyclotron resonance both build upon the fact that masses can be represented as frequencies and frequencies can be measured very accurately. Common for both of these instruments is that the ions are allowed into a chamber, in which a strong magnetic field is present. The shapes of these chambers vary. However, in both the ions start to oscillate back and forth between the axial ends of the chamber, like a pendulum. The frequency of these oscillations is

$$\omega = \sqrt{\frac{e}{m/_{z}} \cdot k}$$

Where k is a constant proportional to the potential difference between the central and the outer electrodes [59].

Thus, all ions with the same m/z ratio will oscillate with the same frequency. It can also be noted that the frequency of these oscillations does not depend upon the initial velocity or coordinates of the ions. The outer electrode of the Orbitrap is split into two. This allows for the detection of the current the ions give rise to as they oscillate in the Orbitrap. For this to happen, it is important that the ions oscillate in a coherent, concentrated packet [59]. This signal is usually measured for about 1 second. Ideally, longer sampling times would be useful as they would provide more accurate measurements, with fewer artefacts present in the spectrum. However, collisions with residual gas in the chamber make this impossible, as it disrupts the ions. After acquiring this spectrum (the current over time), it is converted into an m/z spectra using Fourier transform. Because the sampling time often is not long enough, artefacts are present in the spectra as small peaks. To get a better spectrum, these smaller peaks are often removed, a process called apodization. This removes the false peaks, but it also widens the actual peaks (thus lessening the resolution) [54], [59].

4.6.3 Ion Detection

After separation, the ions are detected. Since the incident ions generate a very weak current, amplification is necessary to achieve a useful signal. This amplification is commonly done by using an electron multiplier. The incident ions will strike a surface, which releases electrons. By applying the correct voltage, these electrons can be forced to strike another surface, releasing more electrons. This is repeated until the amount of electrons is high enough to detect the generated current. Instruments such as FT-ICR and the Orbitrap rely on different means of detection. The ions present in the chamber will oscillate back and forth, giving rise to an image current on two plates close to the chamber. This current can then be transformed into a signal using Fourier transform [54].

4.6.4 Mass spectrometry terms

4.6.4.1 Mean free path

In mass spectrometry instruments, it is important that the ions are able to travel from the ion generator to the detector without interference. An interfering compound, such as gas molecules, can deflect the ions, thus preventing them from reaching the detector. In addition, collisions can introduce fragmentation in the ion. These unwanted reactions make the resulting spectra more complex. There are situations where ion fragmentation is wanted, but this is done under controlled circumstances. See more under Tandem MS.

The mean free path is the path a particle is able to travel without colliding with other particles. A mean free path of 1 m is required in most instruments (ranging up to several hundred kilometres for an Orbitrap). A fairly accurate estimation of the mean free path is

$$L = \frac{0.66}{p}$$

where L is the mean free path (in cm) and p is the pressure (in Pa) [54].

4.6.4.2 Space charge effects

Space charge effects occur when there are too many ions present in the ion trap. The outer ions will "shield" the inner ions, thus affecting the field. The distorted electrical field will lead to a lack of performance and may affect measurements [54].

4.6.4.3 Resolution

The term resolution of a mass spectrometer refers to its ability to separate peaks that lie close together. For example, CO_2 and C_3H_8 will both have the nominal mass of 44 Dalton. However, the actual masses will be 43.98983 and 44.0628, assuming the most abundant isotopes. Using a mass spectrometer with a high enough resolution, these compounds will be present as two separate peaks at their respective m/z ratios [54], [59].

Of the two methods used for determining resolution, the full width at half maximum (FWHM) is the most common. Using this definition, resolution is calculated as

$$R = \frac{M}{\Delta M}$$

Where M is the m/z value of the peak and ΔM is the width of the peak at half its height. A peak at m/z 1000 with a width of m/z 0.1 at half maximum would thus have a resolution of 10,000. The resolution varies between different instruments, from 4000 for a quadrupole to over 100,000 for a modern Orbitrap instrument. The resolution might also vary with the m/z ratio, and how it varies depends on the instrument: [54]

- On quadrupoles and ion traps, the bandwidth is always constant. Thus, if the m/z doubles, the resolution will double with it.
- On TOF instruments and magnetic analysers, the resolution is constant throughout the m/z spectrum. This means that the mass accuracy is good in the lower part of the spectrum, but decreasing as the m/z ratio increases.
- FTICR: the resolution is inversely proportional to the m/z ratio.
- Orbitrap: the resolution is inversely proportional to the square root of the m/z ratio.

4.6.4.4 Mass accuracy

The mass accuracy is the instrument's ability to determine the correct m/z ratio of a molecule. The m/z ratio of a peak is usually calculated by a weighted average of the measurement points around the peak. The mass accuracy is thus intimately coupled to resolution – a better resolution means that nearby peaks have less, or no, influence on the calculation of the m/z ratio of a single peak. However, an accurate estimation of the mass might also be performed on a fairly low-resolution instrument. As long as the sample is pure, containing no other compounds than the analyte and the solvent, the different m/z peaks from the analyte will be isolated [54].

The table below describes some of the properties of the mass spectrometers used in this thesis, as well as other instruments.

	Quadrupole	**Ion trap**	**TOF**	**TOF reflectron**	**Magnetic**	**FTICR**	**Orbitrap**
m/z limit	4000	6000	>1000000	10000	20000	30000	50000
Resolution[*]	2000	4000	5000	20000	100000	500000	100000
Accuracy	100 ppm	100 pm	200 ppm	10 ppm	< 10 ppm	<5 ppm	5 ppm
Ion sampling	Continuous	Pulsed	Pulsed	Pulsed	Continuous	Pulsed	Pulsed
Pressure	10^{-5} Torr	10^{-3} Torr	10^{-6} Torr	10^{-6} Torr	10^{-6} Torr	10^{-10} Torr	10^{-10} Torr

Table 1 - Comparison of Mass Analysers. Selection of data from Table 2.2 in [54].

4.7 Tandem MS

While MS can be used to identify ions from smaller molecules, it is not sufficient for the identification of a peptide from a protein. This is because several peptides might end up with very similar masses. To improve the analysis of larger molecules, the ions may be split into fragment ions. These ions may then be analysed, which gives more information about the original ion (or *precursor ion*). An example of this is that a peptide may be fragmented. The resulting mass spectrum can then give information about the sequence of the peptide, thus allowing for *de novo* sequencing – a technique which has emerged during the last 20 years. However, it might not be necessary to sequence the peptide completely. If the proteome of a sequenced organism is examined, the genome gives information about which proteins should be present in the sample. Thus, it might be sufficient to identify a few peptides to uniquely identify a protein. Apart from access to a sequenced genome, this requires programs able to make the necessary comparisons. This includes analyzing the spectra from the mass spectrometer to identify the sequence and match this with the proteins that should be present. There exists a range of such programs today, such as Sequest and Mascot [51], [54].

4.7.1 Fragmentation

Fragmentation can be carried out in several instruments. One intuitive way is in a triple quadrupole: in the first cell, only ions of a certain m/z ratio, the so-called precursor ions, are selected. This is done by applying a certain combination of RF and DC waves. In the second cell the precursor ion is subject to fragmentation, usually by colliding with neutral gas molecules. In the third and final cell, the product ions are scanned, just as in "single" MS [54].

4.7.1.1 Fragmentation techniques

There exist several ways to fragment peptide ions. Some aspects of fragmentation techniques are the propensity to cleave a certain bond in the backbone and the ability to fragment all such bonds in the peptide, not just bonds N- or C-terminally. Some of the most common techniques of fragmentation are CID, ETD, ECD and HCD and they are described below.

- CID (collision-induced dissociation): with this fragmentation technique, the ions are exposed to an inert gas (commonly He or N_2). The energy added to the ion by the collision will fragment bonds, usually the weakest ones (the Cα-N bond). This is a common technique, but with a certain

[*] Measured using FWHM, at m/z 1000

drawback: it commonly removes PTM’s, thus preventing them from being analysed. The CID method commonly produces b- and y-ions.

- ECD (electron capture dissociation) and ETD (electron transfer dissociation) are similar techniques. Both aim at fragmenting the ionic peptide by collision with electrons, and both mainly produce c- and z-ions. ECD does this by creating free electrons, e.g. from a heated filament, while ETD does this by creating gaseous anions. The anions then transfer one electron to the peptide, which dissociates. ETD is usually better than ECD at retaining PTM’s [61], [62].
- HCD (Higher energy collision dissociation) is a variant of CID, specifically developed for the Orbitrap. The fragmentation is done in a separate chamber, an octopole at the other end of the C-trap and with higher energy [63]. It is supposed to have little to no low cut-off for mass, thus improving the spectra in the low m/z end.

4.7.1.2 Fragmentation of the peptide backbone

Upon fragmentation, the peptide backbone can break in three places between two amino acid residues: Between C-Cα, Cα-N and N-C. For simplicity, it is assumed that the precursor ion only carried one charge. Depending on on which fragment the charge ends up, the product ion will be named differently. If the charge ends up on the N-terminal side of the cleavage site, the ion will be named a, b or c depending on if the cleavage was between C-Cα, Cα-N or N-C, respectively. Similarly, if the charge ends up on the C-terminal side, the ion will be named x, y or z (See figure 3). In addition, the ion will have a number showing how many amino acids are present. So, for a 4-residue peptide, it would be possible to get $b_{1,}$ y_3, b_2, y_2, b_3 and y_1 [54].

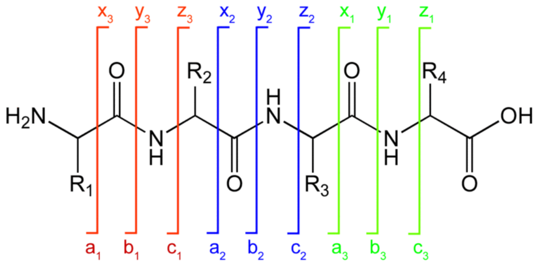


Figure 3 - Peptide fragmentation. Figure adapted from [64]

The method of fragmentation affects the ions that are formed. CID is a so-called ergodic process. This means that the energy spreads out over the molecule, preferentially breaking the weakest bonds. In the backbone, this means the amide bonds which is why b- and y-ions are usually produced with CID. However, PTM’s are also affected, since they usually have weaker bonds. A non-ergodic process is therefore used when PTM’s must be kept for examination [54].

4.7.2 Determining which peptides are present

In principle, it is possible to determine which peptide that was the precursor ion of a certain spectrum by analysing the spectrum manually. This is a tedious process which requires a lot of skill and experience. It might be done if there is only one or a few proteins of interest in the sample, produced e.g. by an upstream biotin-streptavidin affinity chromatography. However, for a large sample such as a digest from a whole cell or more, this would take too much time. Thus, determining which peptides are present is commonly done by employing computers to do the work. Programs such as Mascot or Sequest are used for this.

The program starts with one or more databases of genomic data, such as UniProt’s SwissProt*. The database contains sequences of several hundred thousand proteins. These databases may be curated, and thus of higher quality, or not curated, containing more sequences, possibly duplicates. The user sets

* web.expasy.org/docs/swiss-prot_guideline.html

certain parameters such as the database, the species and the digestive enzyme. Given the genomes in the database, the program virtually digests all the proteins predicted by the genes. For each peptide created this way, the program calculates the spectrum this peptide would give rise to in tandem MS. This requires knowledge about which method of fragmentation that is used. After a run on a mass spectrometer, the acquired spectrum is then compared to the theoretical spectra expected. If an acquired spectrum and a theoretical spectrum are "similar enough", the peptide that gave rise to the theoretical spectra is said to exist in the sample. What "similar enough" means is still discussed [65].

4.7.3 Determination of proteins

In a similar way, the existence of proteins can be inferred from the peptides in the sample. If a protein consists of ten tryptic peptides and nine of these are present in the sample, it is likely that they origin from the protein in question. However, several proteins can give rise to the same sample. Consider the following example.

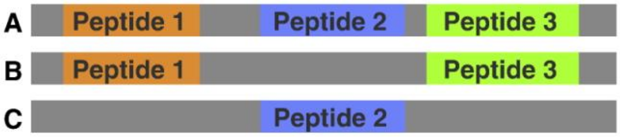


Figure 3: Protein identification. Figure adapted from Cottrell [66]

Let us say that the peptides 1, 2 and 3 are present in the sample. This could indicate either the presence of protein A, or protein B and C together. The principle of parsimony would indicate choosing the simplest explanation – that is, the presence of protein A. However, what would happen if peptide 2 is a very weak, or even a false hit? In that case it would be impossible to determine whether protein A or B is in the original sample. As such, proteins that are present in the original sample might not show up, and proteins that are not present might be present in the analysis. It is common for programs to include a false discovery rate in an analysis, i.e. an estimate of how many of the positive hits that are false [65].

5 Materials and Methods

The general experimental approach in this work was to extract proteins from soil or periphyton samples, followed by purification, digestion and analysis by MS. As a result, steps such as phenol extraction have been used in the preparation of several samples. In this part, I have chosen to describe only the methods themselves, and not the order in which they were used for preparation of samples. Instead, that is described in the Results section.

Appendix 1 contains information on how various buffers and gels were prepared. Since many instruments were used in more than one method, I've listed below which instruments were used for which purposes.

PURPOSE	**INSTRUMENT/EQUIPMENT**
Speed-Vac	Savant SPD1010
Sonication	Branson 1510, Soniprep 150
Spectrophotometer	Beckman DU640
HPLC	Agilent 1100/1200 series
C18 columns	Glygen TopTip 10-200 µl
Mass spectrometry	Bruker HCT Ultra Voyager DE Pro (MALDI-TOF) Orbitrap Velos Pro

5.1 Motivation

Soil samples are highly complex. The microbial communities contained within are likely to be very diverse, since there are no extreme conditions favouring one organism over another. Given the high diversity, soil samples are likely to react to quickly to environmental changes. The periphyton consists of various microorganisms such as algae and cyanobacteria. These are primary producers and, as such, important for the food webs of freshwater and marine environments. In addition, they serve as habitats for other organisms. Organisms present in the periphyton adapt to different ecological conditions caused by physical, chemical and biological disturbances introduced.

5.2 Extraction and Precipitation of Proteins from Soil

5.2.1 Using SDS-phenol

For the soil extraction, a protocol from Keiblinger *et al* [49] was modified. Briefly, the soil was lyophilized. This step was performed by another member of the group and the samples were stored at -20 °C. Extraction buffer (containing Tris, SDS and phenol) was added to 4 g lyophilized soil in a 1:3 (v/v) soil:buffer ratio. The solution was vortexed vigorously for 30 minutes, sonicated for 1 minute on ice and shaken for 1 hour. This procedure was repeated once. The sample was centrifuged at 3,200 g for 20 minutes at 4 °C. The upper phase, containing nucleic acids, was removed. The phenol phase containing proteins was transferred to a new tube, carefully avoiding contamining it with soil particles. To the phenol solution, an equal amount (approximately 5 ml) of H_2O was added. The solution was centrifuged as above. The upper water phase was removed. The phenol phase was transferred to a new tube, carefully avoiding contamination with soil particles.

Soil proteins were precipitated with a 5-fold volume (approximately 30 ml) of 0.1 M ammonium acetate in ice cold methanol and incubated overnight in -18 °C. The tube was centrifuged at 10,000 g for 20 minutes at 4 °C and the supernatant was discarded. The pellet was washed with 2 ml 0.1 M ammonium acetate in ice cold methanol, sonicated, incubated in -18 °C for 15 minutes and centrifuged again as above. The pellet was washed twice with 2 ml cold 80% acetone, resuspended by vortexing, sonicated for about 10 minutes and centrifuged as above. The supernatant was discarded and the sample was dried on the bench for about 10 minutes, to avoid over drying the sample. The pellet was resuspended in 50-100 µl

50 mM NH_4HCO_3. The pH was adjusted by washing with NH_4HCO_3 until neutral, as measured by a pH strip.

5.2.2 Using the SDS-boiling method

A second method for protein extraction from soil was used, which employed boiling of the sample to lyse the cells. 2 g soil was added to a 50 ml Falcon tube. 4 ml alkaline SDS buffer was added. The tube was vortexed vigorously for 3 minutes to disperse the soil. The cells were then lysed by heating the tube to 100 °C for 10 minutes, while shaking at 600 rpm. The slurry was allowed to cool for 5 minutes, after which it was vortexed vigorously for 5 minutes. The tube was centrifuged at 2,100 g for 10 minutes at 4 °C. The supernatant was transferred to a 2 ml Eppendorf tube. The proteins were precipitated by adding 100 % TCA to a final concentration of 25 %, shaking, and incubating at -18 °C overnight. The tube was centrifuged at 20,800 g for 10 minutes at 4 °C. The supernatant was discarded and the pellet was washed with 1 ml -18 °C acetone. This step was repeated three times. The supernatant was discarded and the pellet air-dried for about 10 minutes. The pellet was then dissolved in 2 ml guanidine buffer and incubated at 60 °C for 1 hour at 600 rpm.

5.3 Extraction and Precipitation of Proteins from Periphyton

The periphyton sample was stored in a single bottle in -18 °C. The sample was thawed, aliquoted into pieces with a volume of 1-1.5 ml (dry weight ca 0.5 g) and stored in 2 ml Eppendorf tubes in -18 °C.

5.3.1 Extraction buffers

Three buffers were used for the extraction of proteins from periphyton samples, prep buffer, solubilisation buffer and SDT (SDS-DTT in Tris) buffer. The contents of each is given in Appendix 1.

5.3.2 Mortar Grinding

One sample tube was removed from the freezer and its contents were thawed. The tube was centrifuged at 10,000 g for 5 minutes and the supernatant was discarded. The pellet was placed in a mortar and ground in liquid nitrogen. The lyophilized sample was dissolved in 10 ml prep buffer and homogenized in a Potter-Elvehjem tissue grinder. The sample was sonicated using a Soniprep 150 for 1 minute at an amplitude of 14 micrometers, followed by 9 cycles of 30 seconds sonication-30 seconds “rest” at 18 micrometers. The sample was transferred to a centrifuge tube and centrifuged at 14,000 g at 4 °C for 20 minutes. The supernatant was transferred to a new tube and approximately 40 ml ice cold acetone was added. The tube was shaken a few times and incubated in -18 °C overnight.

5.3.3 Polytron

A tube containing periphyton sample was centrifuged at 10,000 g for 5 minutes and the supernatant was discarded. The sample was dissolved in 10 ml buffer (various buffers were used) and homogenized with the polytron for 10 seconds. It was then homogenized in a Potter-Elvehjem tissue grinder. The sample was sonicated on a Soniprep 150 for 6 cycles of 30 seconds sonication-30 seconds “rest” at 18 micrometers. The sample was transferred to a centrifuge tube and centrifuged at 14,000 g at 4 °C for 20 minutes. The supernatant was transferred to a new tube. The pellet was dissolved in 5 ml 0.0476 M $MgCl_2$ in acetic acid and vortexed for 60 minutes on ice. The sample was centrifuged at 14,000 g at 4 °C for 20 minutes. The supernatant was transferred to a new tube. Approximately 40 ml ice cold acetone was added to the tubes containing the two supernatants. The tubes were shaken a few times and incubated at -18 °C overnight.

5.4 Protein concentration determination

The concentration of protein in extracted samples was determined using the Bradford assay [48]. A reference ladder was prepared with BSA of different concentrations (stock of 0.1 mg/ml diluted 20-100 times) in 200 µl Bradford solution and H_2O to a total volume of 1 ml. The samples were prepared in a similar way. The solutions were incubated for 10 minutes, after which measurements of absorption were performed at 595 nm. If the initial concentration was outside the reference ladder, the sample was diluted and measured again until within range.

5.5 Protein Purification and Cleanup

For protein purification, several methods were employed:

5.5.1 C18 TopTip Columns

The column was conditioned by loading it with 50 µl releasing solution and centrifuged (3,000 g, 2 minutes at room temperature on an Eppendorf Centrifuge 5804R). This was repeated two more times. It was then loaded with 50 µl binding solution and centrifuged as above. This was repeated two more times. The sample was loaded on the column and centrifuged as above. This was stored as flowthrough. The sample was washed by loading the column with 50 µl binding solution. It was then incubated for 2 minutes and centrifuged as above. This was stored as wash. For elution, 50 µl of releasing solution was loaded and the column was centrifuged as above. This was repeated one more time. The flowthrough, wash and eluate fractions were dried on speed-vac and stored in -18 °C.

5.5.2 HPLC

For HPLC, an Agilent 1100/1200 Series was used. The sample was dissolved in 6 µl 5 % ACN, 25 mM KH_2PO_4 and run on a column (Fortis 150x2.1 mm, 5 µm) with a gradient from 25mM KH_2PO_4 5 % ACN to 25 mM KH_2PO_4, 30 % ACN for 60 minutes and the fractions were collected in a 96-well plate. The fractions expected to contain proteins or peptides (as measured by absorbance at 214 nm for peptide backbone and 280 nm for aromatic residues) were pooled together and dried on speed-vac.

5.5.3 Pierce columns

Pierce Detergent Removal Spin Kit from Thermo Scientific, size 0.5 ml, was used for detergent removal. The column was centrifuged at 1,500 g for 1 minute at room temperature. A total of 0.4 ml NH_4HCO_3 was added to the column, followed by centrifugation at 1,500 g for 1 minute. This procedure was repeated two more times. Of the sample, 100 µl was added to the top of the column and incubated for 2 minutes. The column was centrifuged at 1,500 g for 2 minutes and the eluate collected and dried on speed-vac.

5.5.4 FASP protocol

FASP (Filter Aided Sample Preparation), a method where both detergent removal and digestion is performed in the same tube, was used. Millipore Centricon YM-30 columns were used for the preparation.

The protocol was followed with some minor modifications. Briefly, the combination of centrifuge (Eppendorf Centrifuge 5804R) and rotor (A-4-44) used could not create the required forces. Instead of centrifugation at 14,000 g for 10 or 15 minutes, 4,500 g for 20 or 30 minutes was used.

Approximately 20 µg protein in a 1:4 sample:UA (containing urea and Tris) buffer ratio was loaded on the column. For reduction of disulfide bonds, 100 µl UA-DTT was loaded and the column was centrifuged at 4,500 g for 20 minutes. The column was loaded with 100 µl UA and then centrifuged at 4,500 g for 20 minutes. For alkylation, 100 µl UA-IAM was added and the column was shaken at 600 rpm for 1 minute and then incubated in the dark for 20 minutes. The column was centrifuged at 4,500 g for 30 minutes. The column was loaded with 100 µl UA and centrifuged at 4,500 g for 30 minutes. This was repeated twice. The column was then loaded with 100 µl 50 mM NH_4HCO_3 and centrifuged at 4,500 g for 30 minutes. This was repeated twice. Trypsin, (G-Biosciences, sequencing grade, porcine) 0.6 µg in 40 µl 50 mM NH_4HCO_3, was added and the column was shaken at 600 rpm for 1 minute. The column was sealed with parafilm and placed in a box with wet napkins on the bottom. The box was sealed with parafilm and placed in a 37 °C chamber overnight. The filter unit was transferred to a new collection tube and the column was centrifuged at 4,500 g for 30 minutes. The column was loaded with 40 µl 50 mM

NH_4HCO_3 and centrifuged at 4,500 g for 30 minutes. A total of 0.5 µl formic acid was added to the collected elute.

5.5.5 Short SDS-PAGE

Some samples were purified using SDS-PAGE. The proteins were run on the gel for a short time only, to remove detergents and other compounds which might interfere with later analysis. Approximately 25 µg protein was dissolved in 10 µl H_2O. 3 µl sample buffer (4x) was added to the tube. The sample was incubated at 65 °C for 10 minutes, after which it was loaded onto a gel kindly provided by Björn Ingelsson (Department of Clinical and Experimental Medicine, Linköping University). The gel was run for 15 minutes at 62V, 11.9 mA on a gel with 6 % acrylamide, after which approximately 2 cm of the gel was cut into pieces and placed in an Eppendorf tube. The gel pieces were dried in a speed-vac.

5.5.6 SDS-PAGE, mini-gel with Coomassie staining

Minigels (ca 8 cm tall) were cast, with 15 % acrylamide in the resolution gel and 6 % acrylamide in the stacking gel, placed in the machine and covered in running buffer. Volumes corresponding to 20 µg protein were mixed with 20 µl sample buffer (1x). The mixture was incubated at 65 °C for 10 minutes. 15 µl was loaded on the gel. Gels were run at 100 V for stacking and 150 V for resolution. The gels were stained by covering it in Coomassie staining solution and rocking gently for a few hours or until visible protein bands had appeared. The gel was rinsed by covering it in water for 5 minutes, twice, to remove the staining solution. The gel was destained by covering it in destaining solution and rocking gently for a few hours or overnight. The destaining solution was removed by rinsing the gel with water. The gel was stored in water until further use.

5.5.7 SDS-PAGE, long gel with silver staining

For improved resolution, some protein samples were separated on a longer gel and run for a longer time. A 24 cm gel was cast, with 15% acrylamide in the resolution gel and 6% acrylamide in the stacking gel, placed in the machine and covered in running buffer. Approximately 180 µg protein was mixed with 200 µl sample buffer and boiled until completely dissolved, ca 20 minutes. The sample was loaded on a total of 4 lanes. The gel was run for 2 hours for stacking and ca 20 hours for resolution. The gel was fixed overnight in 45% methanol, 45 % H_2O, 10 % acetic acid. The gel was washed in 50 % ethanol in water 3 times for 20 minutes each. The gel was shaken in 200 ml pretreatment solution for 1 minute. The gel was then rinsed in ca 150 ml milliQ H_2O 3 times for 20 seconds each, followed by shaking in 100 ml silver solution (containing silver nitrate) for 20 minutes. The gel was rinsed with ca 150 ml milliQ H_2O and then washed with developing solution until bands were visible. To stop the developing reaction, the gel was incubated with stop solution for 10 minutes. The gel was stored in incubation solution until further use.

5.6 Digestion

5.6.1 In-solution digestion

For reduction of disulfide bonds, 2 µl 100 mM DTT was added to 50 µl sample and incubated for 1h. For alkylation, 6 µl 100 mM IAM was added to the sample and incubated in darkness at room temperature for 30 minutes. Trypsin (sequencing grade, porcine), 3 % of total protein mass, was added and the sample was incubated overnight at 37 °C. The tube was centrifuged for 5 minutes at 13,000 g and the supernatant was transferred to a new tube.

5.6.2 In-gel digestion

The dried gel pieces were dissolved in 400 µl 10 mM DTT in 25 mM NH_4HCO_3 and incubated at 56 °C for 1 hour. The pieces were cooled to room temperature and the excess solution was removed. For alkylation, 400 µl 55 mM IAM in 25 mM NH_4HCO_3 was added and the sample was incubated at room temperature in the dark for 45 minutes. The solution was removed and 500 µl 25 mM NH_4HCO_3 was added. The sample was incubated at room temperature for 10 minutes with occasional vortexing. The solution was removed and replaced with 500 µl 1:1 25 mM NH_4HCO_3:ACN. The sample was incubated at room temperature for 10 minutes with occasional vortexing. This was repeated once. The supernatant was discarded and the gel pieces were dried in speed-vac. The gel pieces were rehydrated in a mixture of 500 µl 25 mM NH_4HCO_3 and 12.5 µl 0.2 mg/ml trypsin and incubated overnight at 37° C. A total of 250 µl H_2O was added to the sample. The sample was incubated at room temperature for 20 minutes with occasional vortexing. The solution, containing peptide extracts, was transferred to a new tube. To extract more peptides, 250 µl extraction buffer (5 % TFA in 1:1 H_2O:ACN) was added to the tube containing the gel pieces. The tube was vortexed occasionally for 20 minutes, after which the solutions containing peptide extracts were pooled. This was repeated once. The tube containing the three pooled extracts was dried in the speed-vac.

5.7 MS Analysis

Most samples were analyzed by MALDI to see if they contained any peptides. Samples were then analysed on the Bruker HCT Ultra or the Orbitrap Velos Pro for a more thorough analysis with higher resolution.

5.7.1 MALDI

A mixture of 0.5 µl sample and 0.5 µl matrix (α-cyano-4-hydroxycinnamic acid in 70 % ACN, 0.3 % TFA) was prepared and placed on a MALDI plate. The mixture was allowed to dry for ca 10 minutes, after which it was inserted into the machine. Spectra were acquired with 300 shots in the 700-7000 Da range, with low pass filtering at 500 Da.

5.7.2 Bruker HCT ultra

The samples were suspended in 11 µl 0.1% FA and spun in Eppendorf centrifuge 5415C for 10 minutes at 13,800 g. 5 µl of the supernatant was loaded on a mass spectrometer vial and placed into the Thermo nanoLC system, followed by injection into the mass spectrometer. On the Bruker HCT Ultra, each sample was followed by flushing the system with blanks:

Type	**LC method**	**MS/MS method**
Sample	(1) or (2)	(5)
Blank	(3)	
Blank	(3)	
Repeat for other samples…	…	…
Blank, 300 min	(4)	

Table 2 - MS Schedule

5.7.2.1 *LC method, 90 min (1)*

Flow rate: 0.3 µl/min
Buffer A: 0.1 % FA in H_2O
Buffer B: 0.1 % FA in ACN

Time (min)	**Concentration B (%)**
0	0
60	35
80	100
90	100

5.7.2.2 *LC method, 240 min (2)*

Flow rate: 0.3 µl/min
Buffer A: 0.1 % FA in H_2O
Buffer B: 0.1 % FA in ACN

Time (min)	**Concentration B (%)**
0	0
120	20

200	40
220	100
240	100

5.7.2.3 LC method for blanks, 30 min (3)

Flow rate: 0.3 µl/min
Buffer A: 0.1 % FA in H_2O
Buffer B: 0.1 % FA in ACN

Time (min)	**Concentration B (%)**
0	0
19	40
25	100
30	100

5.7.2.4 LC method for blanks, 300 min (4)

Flow rate: 0.3 µl/min
Buffer A: 0.1 % FA in H_2O
Buffer B: 0.1 % FA in ACN

Isocratic flow of 97 % buffer A, 3 % buffer B.

5.7.2.5 MS/MS method for analysis of samples (5)

Precursor ions with an m/z ratio within 200-1500 were fragmented using CID. After acquiring 2 identical spectra, precursor ions were automatically excluded for 60 seconds.

5.7.2.6 Data analysis

The database searches were performed according to [49].

The MASCOT Search Engine[*] was used for protein database searches. Data were searched against a database containing all entries in SwissProt, on various dates from March 1st to May 15th. The following search parameters were applied: (i) trypsin was chosen as the protein-digesting enzyme, allowing for up to two missed cleavages and (ii) carbamidomethylation of cysteine was chosen as a fixed modification. Searches were performed with a precursor ion mass tolerance of up to 1.2 Da and a fragment ion mass tolerance of 0.6 Da.

5.7.3 Orbitrap Velos Pro

The samples were suspended in 11 µl 0.1% FA and spun in Eppendorf centrifuge 5415C for 10 minutes at 13,800 g. 5 µl of the supernatant (corresponding to 0.25 µg protein) was loaded on a mass spectrometer vial and placed into the Thermo nanoLC system, followed by injection into the mass spectrometer. On the Orbitrap Velos Pro, each sample was followed by flushing the system with blanks.

5.7.3.1 LC method for samples

After trials, a simple 133-minute method was found to give consistent and good results.
Flow rate: 8 µl/min
Buffer A: 0.1 % FA in H_2O
Buffer B: 0.1 % FA in ACN

Time (min)	**Concentration B (%)**
0	2
2	2
122	90

[*] http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS

130	90
133	2

5.7.3.2 LC method for blanks

Flow rate: 8 µl/min
Buffer A: 0.1 % FA in H_2O
Buffer B: 0.1 % FA in ACN

Time (min)	**Concentration B (%)**
0	2
5	50
22	90
26	90
29	2
30	2

5.7.3.3 MS/MS method for analysis of samples

A first scan was performed in the Orbitrap, with a scan range of 380 – 2,000 Da. The most intense ions from the first step were fragmented in the ion trap using CID. Ions with a charge state of 1 were excluded. The dynamic exclusion was set to 30 seconds.

5.7.3.4 Data analysis

The spectra were used for databases searches using the PEAKS 6.0 Search Engine [67].The following search parameters were used: (i) trypsin was selected as the protein-digesting enzyme and up to three missed cleavages were allowed and (ii) carbamidomethylation of cysteine was chosen as fixed modification. Searches were performed with a parent-ion mass tolerance of ±15.0 ppm and a fragment-ion mass of ±0.5 Da. Searches were performed against a variety of metagenomic databases.

6 Results

6.1 Process analysis

To summarise, the project did not follow the plan laid out initially. Briefly, the mass spectrometry instrumentation caused a lot of headaches and delays, and the analysis of the mass spectrometry data turned out to be much more difficult than initially expected. The thesis work, which was scheduled to finish in June, continued over the summer and finished in September instead, still not reaching the goals set in the plan.

The first steps, however, went relatively fine. The literature study was finished on time. A few additional papers and other sources were found and incorporated throughout the work, but the major part was finished in the first week. For managing the references, the Papers system was chosen because of its integration with Word. The first week of laboratory work, evaluating methods for the preparation of peptides for mass spectrometry analysis, was delayed for a while since the phenol that was ordered didn't arrive on time. As there was also some trouble with the HPLC and the delivery of a new reverse-phase column took time, the separation of peptides was delayed.

The first runs on the mass spectrometer took place in late February, so the delay up until then was not too severe, at least in my eyes. The first runs took place on the Bruker HCT Ultra the department has access to. However, the samples were most likely placed in the wrong position in the sample tray in the nano-LC prior to the mass spectrometer. Thus, no peptides could be identified and no proteins inferred. We came to the conclusion that there were no proteins in the sample. At this point, we should have performed control experiments to test this. Instead, we began evaluating new methods for protein extraction, separation and digestion of proteins for mass spectrometry. This led to the first major delay in the project, since a lot of time was wasted on this. In hindsight, the decision to use the Bruker HCT Ultra for mass spectrometry was somewhat odd, since the department also has access to an Orbitrap Velos Pro – an instrument more suited for handling complex samples.

The first samples were run on the Orbitrap in early June, or about 3 months after the mass spectrometry protocols were initially set to be finished. During this time, the decision had been made to postpone the presentation until the autumn and continue working during the summer, since the Orbitrap runs promised more rewarding results than those achieved so far. The work during the spring had mainly consisted of evaluating various protocols for protein extraction, from both soil and periphyton samples. So, although the project had been severely delayed, the time remaining until the work was to be finished – about 3 months – remained the same. Since the complete method – from protein extraction to analysis on mass spectrometers and protein identification – had not been finished, the decision was taken to continue developing the method further, instead of applying it to the complete set of samples we had access to.

The next step in the process, identification of peptides and proteins from mass spectrometry data, turned out to be more difficult and time-consuming than initially thought, and delayed the project further than the mass spectrometer issue. As shown in this chapter, protein identification from our samples was very low. Solving this in a proper way could only be done by sequencing the organisms present in the sample, which was an impossible task given the time and resources remaining in the project. The way around this was to evaluate different genomic databases, with the hope that the continued the same or similar organisms as in our sample. This turned out to be quite difficult and no good results were achieved, which is why the final step – drawing conclusions from which proteins were present – was not really performed.

6.2 Soil and periphyton samples

Soil is likely to be one of the most complex sample types regarding metaproteomics. The microbial communities contained within soil are usually very diverse, unlike communities at e.g. acid mine drainages. As such, soil samples are likely to react quickly to environmental changes.

The periphyton consists of various microorganisms such as algae and cyanobacteria. These are primary producers and, as such, important for the food webs of freshwater and marine environments. In addition, they serve as habitat for other organisms. Organisms present in the periphyton, such as diatoms, adapt to different ecological conditions caused by physical, chemical and biological disturbances introduced [68].

6.3 Protein extraction and precipitation

The methods used for protein extraction and precipitation seemed to work well for both soil and periphyton samples. The SDS-phenol method extracted an average of 270 µg protein from soil (corresponding to 67.5 µg protein per g soil). The SDS-boiling method extracted an average of 644 µg protein from soil (corresponding to an average of 161 µg protein per g soil). The extraction of protein from periphyton samples was must successful using polytron and extraction buffer (an average of 167 µg, corresponding to 418 µg protein per g periphyton).

6.4 Protein separation

6.4.1 HPLC

HPLC was performed with peptides derived from a soil sample, extracted using SDS-phenol and digested with trypsin. Figure 4 displays the chromatogram, collected on an Agilent 1100/1200 series at 214.4 and 280.4 nm.

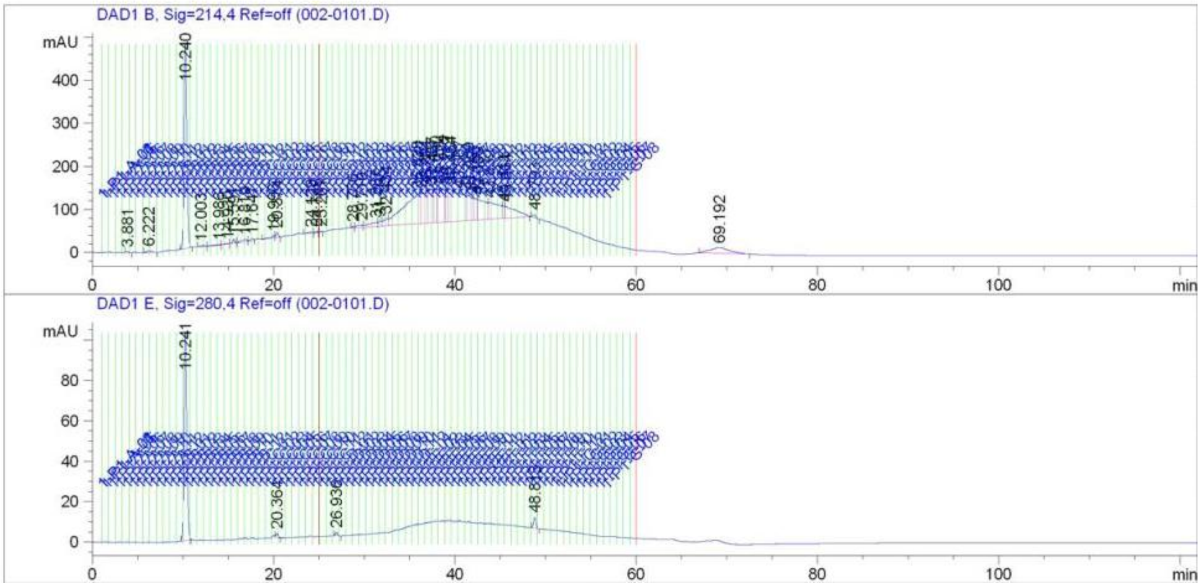


Figure 4 - Chromatogram (measured at 214.4 and 280.4 nm) of soil proteins, digested with trypsin and purified on C18 Toptips.

6.4.2 SDS-PAGE

Protein separation on SDS-PAGE was carried out in several ways on both types of samples. As a general rule, Coomassie-stained gels resulted in long smears, with no individual bands. Silver staining, being more sensitive, stained the gels so that individual bands could be detected. Figures 5 and 6 show typical gels after separation:

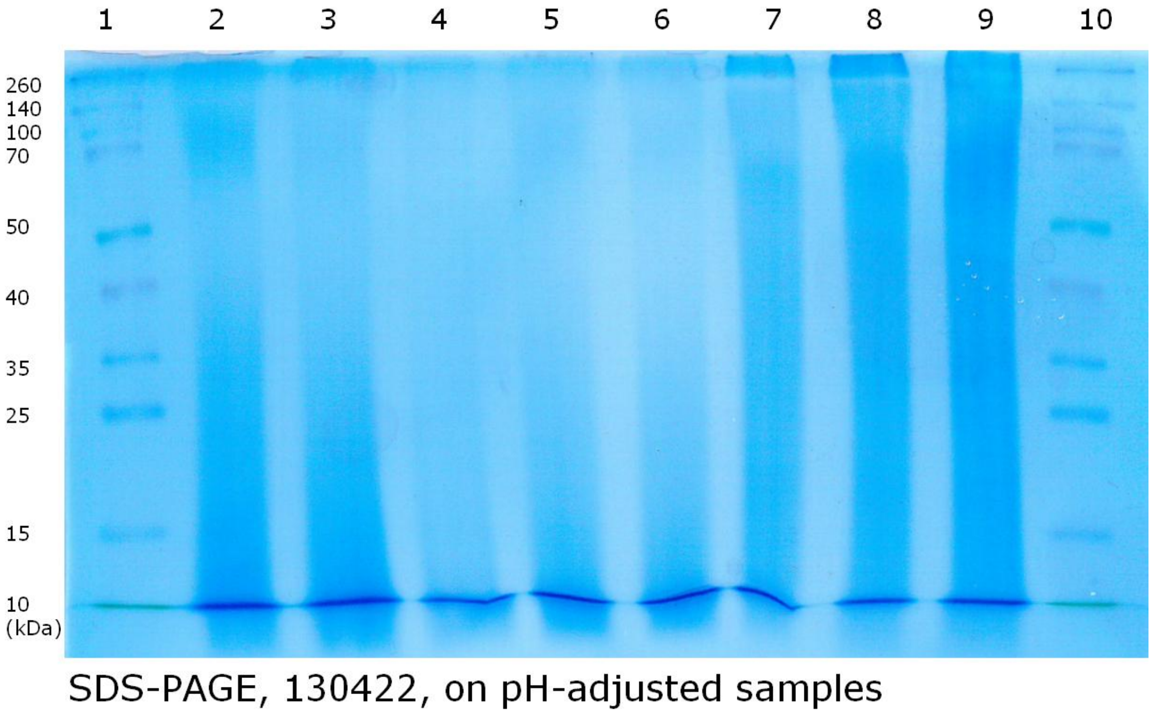


Figure 5: SDS-PAGE on soil and periphyton proteins, stained with Coomassie

Figure 5 depicts the resolving part of an SDS-PAGE, 15 % acrylamide, run at 100 V for 30 minutes for stacking and 150 V for 75 minutes for resolving, then stained using Coomassie solution. Several gels were prepared like this. They all had the same appearance (long smears) and the same result (no protein identification using either the MALDI-TOF or the Bruker HCT Ultra) so this gel is considered representative for all gels stained with Coomassie.

The lanes are the following:

Lane	**Type**
1	Ladder (Thermo Spectra Multicolor, 10-260 kDa)
2	Periphyton sample, extracted with mortar grinding, pH-adjusted with NH_4HCO_3 until neutral
3	Periphyton sample, extracted with mortar grinding
4	Periphyton sample, extracted with mortar grinding, stored in fridge for a few days
5	Periphyton sample, extracted with polytron, pH-adjusted with NH_4HCO_3 until neutral
6	Periphyton sample, extracted with polytron
7	Soil sample, extracted with SDS-phenol
8	Soil sample, extracted with SDS-phenol, pH-adjusted with NH_4HCO_3 until neutral
9	Soil sample, extracted with SDS-phenol, stored in fridge for a few days
10	Ladder (Thermo Spectra Multicolor, 10-260 kDa)

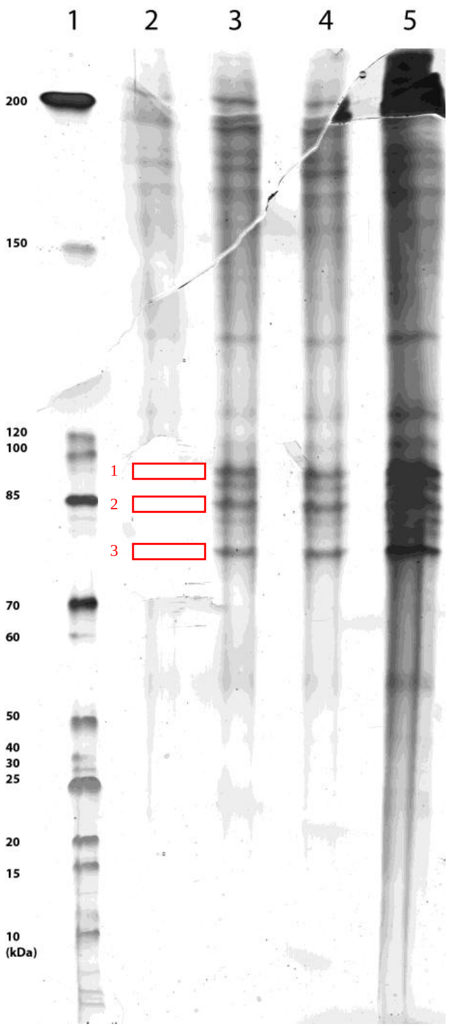


Figure 6: SDS-PAGE on periphyton samples, stained with silver

Figure 6 depicts the resolving part of a long SDS-gel (24 cm) run for ca 20 hours and stained with silver. The left lane is the ladder (Fermentas PageRuler Unstained Protein Ladder, 10-200 kDa), while the remaining lanes are the sample (periphyton sample, proteins from the acetic acid extraction). The sample on the right was loaded first, before noting that it was not completely dissolved. The rest of the sample was thus boiled and vortexed before loaded on the three lanes in the middle, explaining the difference in appearance between the four sample lanes. This gel was scanned after three bands were cut out. The appearance of the three bands that were cut was very similar to the corresponding bands in lane 3.

6.5 Protein Identification

6.5.1 From samples separated on HPLC

The fractions containing the peaks in the chromatogram were pooled. The aim was to perform analysis of these fractions on Bruker HCT Ultra. Unfortunately, it is likely that these samples were never analysed due to a handling error. It is therefore impossible to tell whether the use of HPLC for pre-analysis fractionation was a successful method for the identification of proteins.

6.5.2 From samples separated on gels

Three bands from lane 5 and 8 from the gel in Figure 5 were cut out, digested, and purified. The peptides were analysed on MALDI, but no peptide peaks could be seen. Excised bands from another gel, with samples prepared in a similar way, were analysed on the Bruker HCT Ultra with no protein identification apart from common contaminants. This is why we chose not to carry on with this type of separation. It is likely though that analysing the bands on the Orbitrap would result in protein idenfication.

Bands 1, 2 and 3 from the gel in Figure 6 were cut out, digested with trypsin as described, purified on C18 columns and analysed with MS/MS on the Orbitrap Velos Pro. Acquired spectra were matched against the SwissProt database using PEAKS 6.0. Below are tables containing the five proteins from each band with the highest score in PEAKS. If the same protein was identified in several species, only the first hit was reported.

Accession	-10lgP	Coverage (%)	#Peptides (Unique)	Description
P53476|ACT_TOXGO	146.09	11	2(1)	Actin OS=Toxoplasma gondii GN=ACT1 PE=3 SV=1
P21219|NHA1_RHORH	117.44	11	2(2)	High-molecular weight cobalt-containing nitrile hydratase subunit alpha OS=Rhodococcus rhodochrous GN=nhhA PE=1 SV=3
P80737|H41_BLEJA	72.20	10	1(1)	Histone H4-1 (Fragment) OS=Blepharisma japonicum PE=1 SV=3
P24163|G3P_ENTAE	67.97	2	1(1)	Glyceraldehyde-3-phosphate dehydrogenase (Fragment) OS=Enterobacter aerogenes GN=gap PE=3 SV=1
Q07051|EF1A_EIMBO	67.46	3	1(1)	Elongation factor 1-alpha (Fragment) OS=Eimeria bovis PE=2 SV=1

Table 1 – Top five unique proteins identified in band 1.

Accession	-10lgP	Coverage (%)	#Peptides (Unique)	Description
P21219|NHA1_RHORH	102.79	11	2(2)	High-molecular weight cobalt-containing nitrile hydratase subunit alpha OS=Rhodococcus rhodochrous GN=nhhA PE=1 SV=3
Q31UQ6|KGUA_SHIBS	101.22	13	2(2)	Guanylate kinase OS=Shigella boydii serotype 4 (strain Sb227) GN=gmk PE=3 SV=1
P53476|ACT_TOXGO	85.85	5	1(1)	Actin OS=Toxoplasma gondii GN=ACT1 PE=3 SV=1
A9BP42|UBIB_DELAS	55.58	2	1(1)	Probable ubiquinone biosynthesis protein UbiB OS=Delftia acidovorans (strain DSM 14801 / SPH-1) GN=ubiB PE=3 SV=1
Q98QV2|IF3_MYCPU	49.68	5	1(1)	Translation initiation factor IF-3 OS=Mycoplasma pulmonis (strain UAB CTIP) GN=infC PE=3 SV=2

Table 2 – Top five unique proteins identified in band 2.

Accession	-10lgP	Coverage (%)	#Peptides (Unique)	Description
P21219|NHA1_RHORH	100.70	11	2(2)	High-molecular weight cobalt-containing nitrile hydratase subunit alpha OS=Rhodococcus rhodochrous GN=nhhA PE=1 SV=3
P53476|ACT_TOXGO	85.77	5	1(1)	Actin OS=Toxoplasma gondii GN=ACT1 PE=3 SV=1
P24751|G3P1_ESCVU	56.21	2	1(1)	Glyceraldehyde-3-phosphate dehydrogenase (Fragment) OS=Escherichia vulneris GN=gap PE=3 SV=1
Q8E6F5|Y613_STRA3	53.01	2	1(1)	Uncharacterized RNA methyltransferase gbs0613 OS=Streptococcus agalactiae serotype III (strain NEM316) GN=gbs0613 PE=3 SV=1
O67161|G3P_AQUAE	50.98	4	1(1)	Glyceraldehyde-3-phosphate dehydrogenase OS=Aquifex aeolicus (strain VF5) GN=gap PE=1 SV=1

Table 3 – Top five unique proteins identified in band 3.

Several proteins are present in all three bands. It is a known fact that the same protein can end up in different spots on a gel. This might depend on strain variations, post-translational modifications or issues arising during the analysis of the data [69]. See the following chapter for a discussion regarding these protein hits.

6.5.3 From non-separated samples

Several samples, from whole-solution digests to excised gel parts, were analysed on MALDI. Likely due to the high complexity of the samples, all spectra are fairly similar. Figure 7 shows a representative spectrum.

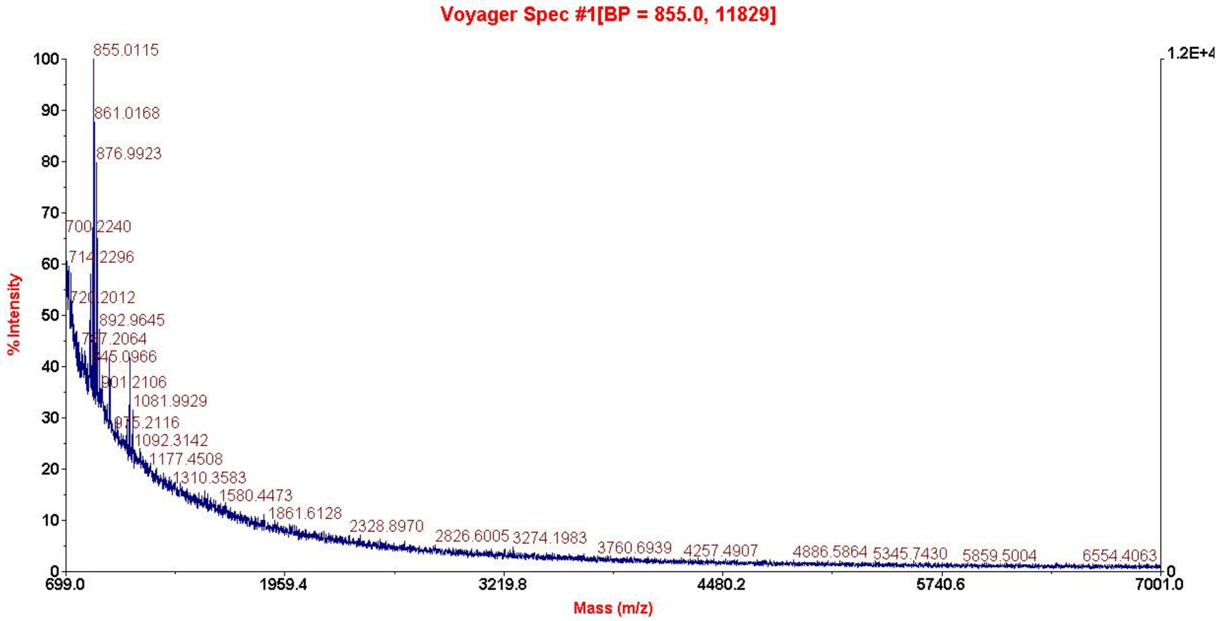


Figure 7 - A typical MALDI spectrum

The sample was prepared from a periphyton sample in preparation buffer run on SDS-PAGE, in-gel digested and purified on C18 TopTip. A few distinct peaks are noticeable. However, this is not enough to identify proteins. Spectra from other samples are very similar and thus not shown.

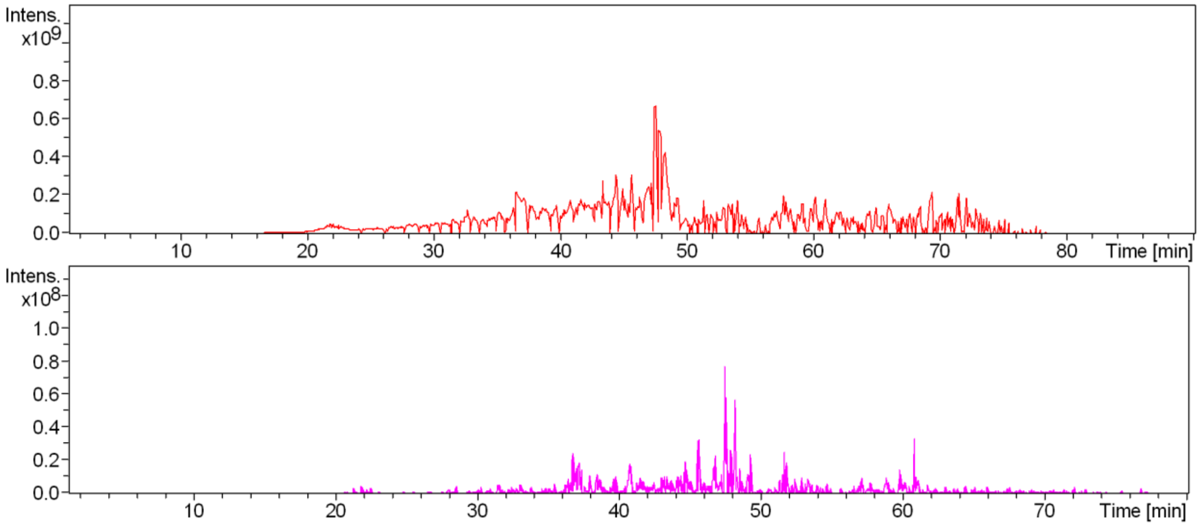


Figure 8 - Typical spectra from the Bruker HCT Ultra

Figure 8 shows typical chromatograms (intensity measured using TIC) for MS (top) and MS/MS (bottom) acquired on the Bruker HCT Ultra. These spectra are from the same sample as the one analysed on MALDI above.

Samples were also analysed using the 240 minutes method described earlier, with the hypothesis that better separation would improve identification. The chromatograms were similar as well as the search results. The spectra from the above MS/MS runs were searched against SwissProt and NCBInr using the free Mascot MS/MS ion search. Removing common contaminants such as keratin from the search result leaves no protein identification.

6.5.3.1 Orbitrap Velos Pro

PEAKS searches revealed proteins in periphyton (both in the whole sample and in the bands excised from the long gel) as well as in the soil. The MS/MS spectra from a periphyton sample run on the Orbitrap were searched against four different databases, in order to investigate which one was best suited for searching with this type of samples. These databases were UniProt/SwissProt (UniProts curated database), UniMES (metagenomic data from Craig Venter's Global Ocean Sampling Expedition), metagenomic data from Yellowstone Lake and data from a Japanese study investigating the bacterial diversity in a river used to produce drinking water.

The sample used for these runs was a periphyton sample, prepared using the solubilisation buffer. A total of 3807 MS spectra and 15413 MS/MS spectra were acquired.

The table below shows the number of identified proteins from each database.

Database	Protein identifications
SwissProt[*]	116
UniMES[†]	164
Yellowstone lake[‡]	10
Japanese river study[§]	1

SwissProt and UniMES were the only two which identified a reasonable number of proteins from the spectra. The majority of proteins in the UniMES database are still uncharacterized and we decided to use SwissProt for further searches because of its superior annotation.

A comparison of the five methods for extraction of proteins from periphyton samples was made. The search was performed against SwissProt using PEAKS 6.0 with search settings as described in section 5.7.3.4. The FDR was set to 1 % and at least 1 unique peptide was required for the positive identification of a protein. In addition, the required scores for peptide hits were higher than in the search above, explaining the fewer number of hits. If the same protein was identified in several species, only the first hit was reported.

The table below shows the number of identified proteins from each extraction method, with the searches made against SwissProt. No protein was identified in more than one extraction method. See Appendix 2 for the identified proteins.

Extraction method	# MS spectra	# MS/MS spectra	Protein identifications
Prep buffer	3631	6528	52
Solubilisation buffer	6499	16345	29
Solubilisation – acetic acid	8961	7577	3
SDT buffer	9711	4876	3
SDT – acetic acid	8452	11088	0

6.5.3.2 De novo searches

PEAKS 6.0 comes with the ability to perform *de novo* sequencing, i.e. automatic annotation of an MS/MS spectrum to identify the peptide giving rise to the spectrum. This is commonly employed as a part of the

[*] Included in PEAKS 6.0

[†] FASTA file downloaded from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/unimes/unimes.fasta.gz on 2013-07-01

[‡] FASTA file downloaded from ftp://portal.camera.calit2.net/ftp-links/cam_datasets/projects/read/CAM_PROJ_YLake.read.fa.gz on 2013-07-01

[§] FASTA file downloaded from ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA072/SRA072293/SRX270944/SRR833230.fastq.bz2 on 2013-07-01 normal database search, to improve the accuracy and number of hits. It is also possible to do a standalone *de novo* search. This was performed on a sample prepared from periphyton proteins extracted with the solubilisation buffer. Identified peptides were subject to biodiversity analysis using the Unipept tool [14]. Unipept searches the UniProt databases (SwissProt and trEMBL) and calculates, for each peptide, the lowest common ancestor (LCA) in which the peptide is present. From the sample, 2,164 peptides were sequenced de novo. 442 of these were identified, with 298 peptides having the top level ("organism") as its LCA. Most of the remaining peptides belonged to bacteria, with Proteobacteria and Firmicutes being most common.

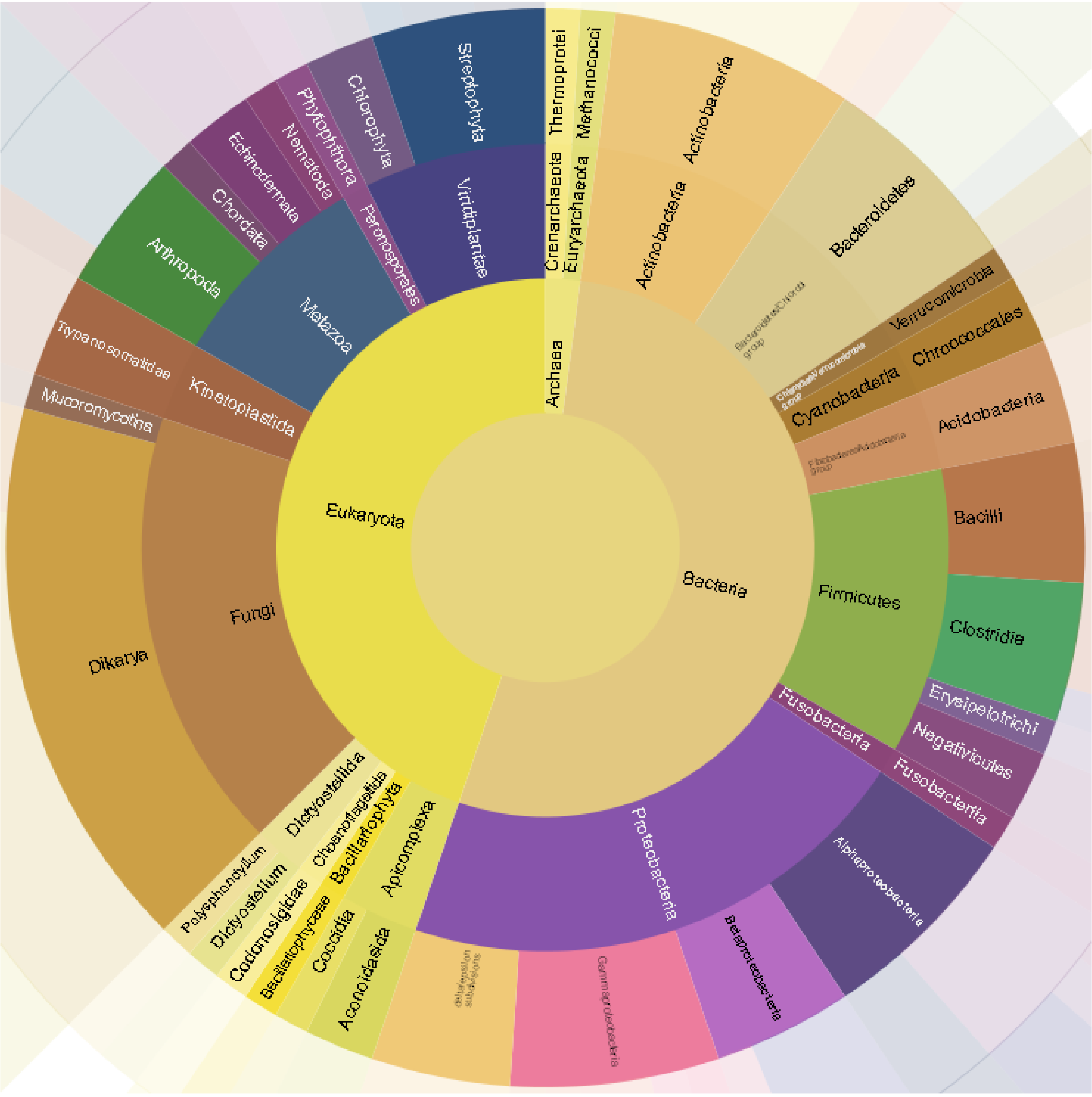


Figure 9 - Sunburst presentation of the 442 peptides that could be matched to an LCA, using Unipept.

7 Discussion: Problem analysis

The protein extraction was satisfactory. The amount extracted was more than enough required for subsequent analysis, with no more than 0.25-1 µg required for analysis on the Orbitrap. An initial concern was whether the proteins would be extracted in proportion to their presence in the original sample, to make any kind of quantification meaningful. This has not been answered in this work, and it is not likely that the chosen methods for extraction are completely unbiased[*]. Another concern is that the treatment of samples prior to protein extraction has affected the proteins that are extracted. To preserve the soil until analysis, it was lyophilized. The soil samples were first intended for a metagenomic analysis, but this was changed to a metaproteomic analysis before this master thesis work began. It is not known whether any previous treatment of the samples affected the proteins. According to Murphy et al, handling proteomic samples on dry ice (which is common in proteomic research and has been used in this work) can cause protein instability and aggregation due to CO_2-induced lowering of pH [70]. Since the soil was lyophilized, thus having a very low content of water, it is unlikely that it was affected by the handling. The periphyton samples, however, being stored in water and shipped on dry ice, might have been affected by this.

A major setback was that many protein extraction methods were never evaluated properly. Many soil samples, digested and purified with a variety of methods, were placed in the wrong spot on the nanoLC-tray upstreams of the Bruker HCT Ultra. Thus, the samples were never analysed. This mistake was not discovered immediately, as we did not quite know what results to expect, and led us to dismiss these methods as not working too soon. It would be of interest to re-run these samples, but perform the mass spectrometry analysis on the Orbitrap for a proper comparison.

The lack of proper metagenomic databases seems to be a major limiting factor. The top five proteins identified in the silver-stained 1D-gel are shown in Table 1-3 in the Results part. These searches were all made against UniProt/SwissProt. The identified proteins are presented together with the species they belong to. Many of the identified species are unlikely to be found in freshwater periphyton samples, from a Spanish river. Of the identified species, several do not exist in this part of the world (*Shigella Boydii* is only found on the Indian subcontinent [71]), prefer other environments (*Aquifex aeolicus* grows at much higher temperatures [72]) or require a host (*Toxoplasma gondii* is a parasite infecting warm-blooded animals [73]). Despite this fact, the protein identification can still be correct. The identified proteins might belong to a related species that has not yet been sequenced, and important proteins can be conserved over long evolutionary distances.

One highly conserved protein is actin. It is conserved over a wide variety of eukaryotic species, with a high level of sequence similarity even between very diverse species. Actin is one of the most common proteins and fulfils several roles, including forming microfilaments for the cytoskeleton. It is thus likely to be present in the periphyton sample (see Table 1-3 in Results), even though it may not originate from *Toxoplasma gondii* [74]. Histones also belong to the group of well-conserved proteins, with histone H4 being the most conserved (likely due to extensive contact with other histones). Histones are proteins responsible for packing the DNA and are found in almost all eukaryotic cells [75]. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) is part of the glycolytic pathway. Its sequence has also been shown to be highly conserved [76].

7.1 Improvements

Below follows a discussion on possible error sources, which could have been avoided to improve the results.

[*] An unbiased extraction method extracts protein proportionally to their presence in the sample

Extracted, precipitated proteins have not always been handled on ice when appropriate. Handling on ice or at 4 °C is common when working with proteins, to prevent proteolytic activity.

However, our approach was similar to other's work where handling on ice is not mentioned [9], [45]. It is therefore likely that increasing handling on ice will only contribute a little to improved results.

Keiblinger *et al* [49] compared four different protein extraction protocols to study their efficiency for two different soil types (forest soil and potting soil). A remarkable result was that out of the 1,982 (forest soil) and 856 (potting soil) unique proteins identified, only 59 and 7 were found with all four methods. Also, out of the 126 versus 326 proteins found in potting plant by NaOH- versus SDS-extraction, only one was found by both. The authors did not come to a conclusion as to why the protocols differed so much regarding the proteins extracted from the soil. It is likely that the different methods of extraction used in this work were similar in that they do not extract the same proteins.

SDS can interfere with the Bradford assay even at low concentrations, by binding to the Coomassie dye. This would give inflated readings of protein concentration, with the subsequent risk of loading too little protein on downstream SDS-PAGE or LC-MS. To the best of our knowledge, we tried to remove the detergent by the precipitation, but it is possible that there was still some left in the sample interfering with the assay [77].

After extraction, the proteins were precipitated. The pellet after this precipitation was not white, but usually coloured brown, green or black depending on what sample it came from. This indicated that there were substances in addition to the proteins in the pellet, since a protein pellet usually is white. It is not known what these substances were, but it might be possible that they interfered with the digestion or the purification. It is also possible that humus, known to have a smearing effect on gels, can affect the identification of proteins [10].

The capacity for a mini gel for mixed proteins is 20-40 µg per well [78], which was approximately the amount that has been loaded on the gels in this work. The capacity is less if a single protein is loaded, since one band will contain the protein molecules. As the samples are highly complex, only very few proteins are expected to be present in high amounts. It is possible that much more proteins need to be loaded on the gel to get enough peptides for analysis on MS.

One major limiting factor of this work was the lack of access to proper metagenomic databases. As stated in the feature article by Hettich *et al* (2013), the best identifications from a metaproteomic sample requires having the metagenome for the same sample. In case the metagenome is not available, related metagenomes as well as synthetic metagenomes can also work [9]. In this work, three metagenomic databases were used. The databases originated from the Global Ocean Sampling Expedition, a Yellowstone freshwater lake and a river used for the production of drinking water. None of these seemed to be related enough to the periphyton samples, as they all provided less information than SwissProt.

8 Conclusions

The work described in this thesis had the aim to extract proteins from different sources, trying to find a method that reliably and consistently extracted proteins in proportion to their presence in the original sample. Proteins were extracted from both soil and periphyton samples, using several methods. For the soil samples, a mixture of SDS and phenol combined with vortexing and sonication proved to work. For the periphyton samples, extracting proteins using the polytron, with the sample in prep buffer, gave the highest yield. All of the methods examined did, however, give yields that were high enough for further treatment – separation, digestion and purification – before subsequent analysis. As such, the prep or SDT buffers are recommended since they are easiest to prepare. The SDS-phenol buffer used for protein extraction from soil can also be prepared relatively fast.

Methods such as FASP, Pierce columns and HPLC were initially dismissed because no proteins could be identified. Initially, we misplaced the vials in the Bruker HCT Ultra, resulting in the samples never being run. After correcting this, the samples were only analysed on the Bruker HCT Ultra for a long time instead of the Orbitrap Velos Pro. Since the Orbitrap Velos Pro has a higher resolution and sensitivity, it is much more likely that proteins will be identified when analysed on this spectrometer. This has also been proven with later samples we decided to analyse on the Orbitrap. It is therefore likely that the methods FASP, Pierce columns and HPLC do work and could be useful for detergent removal and separation. A re-examination of samples prepared with this method on the Orbitrap would be of interest.

All results indicate that the proper detection of complex samples such as these require an instrument with as high sensitivity and resolution as possible. It is therefore recommended to go straight to the most advanced instrumentation as soon as possible. The alternative is to separate the proteins or peptides to a much higher degree than described in this thesis, to decrease the complexity enough for analysis on e.g. the Bruker HCT Ultra. It is likely that this would be both more time-consuming and expensive than the methods outlined in this thesis.

For the data analysis, there is a lot of improvement that can be made. First of all, the work would benefit from searching against a proper database – a database consisting of all the organisms found in the sample. This can be done by sequencing all the organisms from a soil or periphyton sample. This is in itself a massive amount of work and would probably require at least another Master Thesis' work. In the meantime, databases likely to contain organisms found in the samples can be used. Our work, however, showed that such databases (the Yellowstone freshwater database, the Japanese river database and uniMES) were not helpful in identifying proteins. In fact, they performed worse than SwissProt. Another drawback with such metagenomic databases is that most of them are currently unannotated – in other words, the function and species of identified protein sequences are unknown. This can to some extent be remedied with homology searches, but it is not optimal. Unannotated databases give little to no information about the identified proteins' roles. This means that cellular pathways cannot be identified, giving no clue as to how a sample reacts to environmental stress. A well-annotated, relevant metagenomic database would be optimal for a project like this.

For the protein identification, it seems that the extracted proteins vary between the extraction methods. This is likely to happen, and is also seen in the paper by Keiblinger *et al* [49]. To identify as many proteins as possible, it is probably best to employ a mixture of extraction methods and solutions, then perform the database search on the combined spectra from all extractions.

The proteins found in the periphyton sample run on the silver-stained gel (Table 1-3 in Results) all originate from species unlikely to be found in such a sample. Most of the proteins that were identified are highly conserved among a wide variety of samples, likely because of their high importance to survival. It is therefore probable that these identifications are in fact correct, but that the proteins originate from other species (with the same or highly similar protein sequence). The ability to identify these depend on the fact that they are so highly conserved. This likely means that less-conserved proteins, belonging to species that are not yet sequenced and entered into Swissprot, will not be identified. Highly conserved proteins are often important for various parts of a cell's survival, and are expressed at a fairly regular level [79].

Given the last years' improvements in sequencing and information technology, it is likely that the increase in metagenomic databases will and the databases will probably become better annotated over time. This leads to the conclusion that the work described in this thesis will probably provide a viable method for assessing environmental status eventually. The work required to perform these analyses will likely decrease as the threshold for data analysis becomes lower – performing database searches today requires less manual labour than before.

9 Future Prospects

The development of this type of metaproteomic profiles would provide a well-defined protocol to allow adequate comparison of results. The results of this pilot experiment will be the basis to apply a function-based comparative environmental proteomics approach on a larger scale project that could link environmental stressors (pollution, climate changes, etc) to changes in phylogenetic diversity.

- Susana Cristobal, Master Thesis Proposal

The methods outlined in this thesis were effective in extracting proteins from environmental sources. However, the subsequent analysis by mass spectrometry failed to provide an adequate amount of protein identifications. Considering that others have had success with similar projects, we believe it to be possible with the material we have examined as well. It would therefore be interesting to study protocols for protein extraction from other samples. For example, Keiblinger *et al* (2012) have tested protein extraction from soil with SDS as well as with sodium hydroxide, with protein identification from both extraction methods [49].

It is also possible that more protein is needed for a positive identification, since every protein is present in very low amounts. It is possible that a method with few steps can give a higher amount of proteins, since every step in the process is likely to cause sample losses. In their 2013 article, Ning *et al* employ a one-stop workflow for protein extraction using amphipols [80]. They perform protein extraction from a cell culture, but with some modifications the protocol can likely be applied to more complex environmental samples as well.

Environmental samples are highly diverse, and soil samples likely belong to the most complex ones. It is not certain that a protocol that successfully extracts protein from one sample will automatically extract proteins from another sample. One of the protocols used in this work was adapted with modifications from Keiblinger *et al* [49]. It would be interesting to see how the protocols used in this work perform on samples from other types of soil.

An important factor to take into consideration is that a good separation of proteins or peptides before mass spectrometry often increases the amount of proteins detected. This is because a signal from a peptide present in very low amounts might be lost in the signal from other peptides. Gel electrophoresis can separate a complex sample into individual proteins. This level of separation is not quite necessary for the type of work described in this thesis. In addition, it is more time-consuming than online systems such as LC. Improving separation of peptides by introducing further LC separation steps would likely improve detection of peptides. One possible way is to separate the peptides with HPLC. Another way is to introduce a second dimension of separation right before LC-ESI, as described in Erickson *et al* [26].

Due to the lack of success with protein identification, it was impossible to perform an analysis on the extracted proteins. However, if future studies provide methods for protein identification, this is a highly interesting area. The group has access to soil samples that have been treated in different ways, ranging from no treatment to exposure to propranolol and high levels of salinity. A statistical analysis could thus be performed and comparisons made between the different types of samples. Such a comparison could study which proteins are over- or underexpressed or if some proteins get different post-translational modifications. On a higher level, this could be used to infer if some organisms are present at different levels in the samples. A possible result of this is that sentinel organisms can be identified – organisms which can be used to assess the status of an environmental sample on a different, more detailed level than before.

10 Acknowledgments

I would like to thank the following persons:

Susana Cristobal at IKE, for providing me with the opportunity to do my master thesis here as well as supervising it.

Karin Enander at IFM, for the course that got me interested in proteomics, connecting me to Susana and examining this thesis.

Hanna Klang Årstrand at IKE, for aiding me with the protocols, introducing me to the labs and giving me countless tips.

Everyone in Susana’s group for teaching me how various equipment and instruments worked, as well as giving me a warm welcome to the lab.

Everyone else at floor 13 for input, socialization and interesting discussions.

Maria Turkina and **Stefan Ljunggren** for running the samples on the Orbitrap and introducing me to the tools for the computer analysis.

Helena Malmrot, the opponent of this thesis, for valuable discussions.

My friends and family, for pushing me when things didn’t go as planned and cheering me up when they did.

This thesis concludes my studies at Linköping University. It has been a great time.

Henric Fröberg
Linköping, September 2013

11 Bibliography

[1]H. J. Albering, S. M. van Leusen, E. J. Moonen, J. A. Hoogewerff, and J. C. Kleinjans, “Human health risk assessment: A case study involving heavy metal soil contamination after the flooding of the river Meuse during the winter of 1993-1994.,” *Environmental health perspectives*, vol. 107, no. 1, pp. 37-43, Jan. 1999.

[2]P. Brookes, “The use of microbial parameters in monitoring soil pollution by heavy metals,” *Biology and Fertility of soils*, Jan. 1995.

[3]P. Verlicchi, M. al Aukidy, and E. Zambello, “Occurrence of pharmaceutical compounds in urban wastewater: removal, mass load and environmental risk after a secondary treatment--a review,” *The Science of the total environment*, vol. 429, pp. 123-55, Jul. 2012.

[4]B. Schumacher, (2013, October 7) “*Guidance Document for Soil-Gas Surveying*,” [Online]. Available: http://www.epa.gov/esd/factsheets/soil-gas.pdf

[5]D. J. Hoffman, C. P. Rice, and T. J. Kubiak, “PCBs and dioxins in birds,” in *Environmental Contaminants in Wildlife: Interpreting Tissue Concentrations*, W. Beyer, G. Heinz, and A. Redmon-Norwood, Eds. Lewis Publishers, pp. 165-207.

[6]J. C. Wooley, A. Godzik, and I. Friedberg, “A primer on metagenomics.,” *PLoS computational biology*, vol. 6, no. 2, Feb. 2010.

[7]N. M. van Straalen and M. E. Feder, “Ecological and evolutionary functional genomics--how can it contribute to the risk assessment of chemicals?,” *Environmental science & technology*, vol. 46, no. 1, pp. 3-9, Jan. 2012.

[8]D. Wu, M. Wu, A. Halpern, D. B. Rusch, S. Yooseph, M. Frazier, J. C. Venter, and J. A. Eisen, “Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees.”, *PloS one*, vol. 6, no. 3, Jan. 2011.

[9]R. L. Hettich, C. Pan, K. Chourey, and R. J. Giannone, “Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities.”, *Analytical chemistry*, vol. 85, no. 9, pp. 4203-14, May 2013.

[10]E. B. Taylor and M. A. Williams, “Microbial protein in soil: influence of extraction method and C amendment on extraction and recovery.”, *Microbial ecology*, vol. 59, no. 2, pp. 390-9, Feb. 2010.

[11]W. P. Blackstock and M. P. Weir, “Proteomics: quantitative and physical mapping of cellular proteins”, *Trends in biotechnology*, vol. 17, no. 3, pp. 121-7, Mar. 1999.

[12]I. C. Lawrance, B. Klopcic, and V. C. Wasinger, “Proteomics: an overview.”, *Inflammatory bowel diseases*, vol. 11, no. 10, pp. 927-36, Oct. 2005.

[13]E. Nägele, M. Vollmer, P. Hörth, and C. Vad, “2D-LC/MS techniques for the identification of proteins in highly complex mixtures.”, *Expert review of proteomics*, vol. 1, no. 1, pp. 37-46, Jun. 2004.

[14]B. Mesuere, B. Devreese, G. Debyser, M. Aerts, P. Vandamme, and P. Dawyndt, “Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples.”, *Journal of proteome research*, vol. 11, no. 12, pp. 5773-80, Dec. 2012.

[15]P. Wilmes and P. L. Bond, “Metaproteomics: studying functional gene expression in microbial ecosystems.”, *Trends in microbiology*, vol. 14, no. 2, pp. 92-7, Feb. 2006.

[16]R. Twyman, *Principles of Proteomics*, 1st ed. Abingdon: Garland Science/BIOS Scientific Publishers.

[17]P. Wilmes and P. L. Bond, “The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms.”, *Environmental microbiology*, vol. 6, no. 9, pp. 911-20, Sep. 2004.

[18]T. Schneider and K. Riedel, “Environmental proteomics: analysis of structure and function of microbial communities.”, *Proteomics*, vol. 10, no. 4, pp. 785-98, Feb. 2010.

[19]B. D. Dill, J. C. Young, P. A. Carey, and N. VerBerkmoes, *Environmental molecular microbiology*, 1st ed. Norfolk, UK: Caister Academic Press, p. 231.

[20]R. Daniel, “The metagenomics of soil.”, *Nature reviews. Microbiology*, vol. 3, no. 6, pp. 470-8, Jun. 2005.

[21]A. Siggins, E. Gunnigle, and F. Abram, “Exploring mixed microbial community functioning: recent advances in metaproteomics.”, *FEMS microbiology ecology*, vol. 80, no. 2, pp. 265-80, May 2012.
[22]R. J. Ram, N. C. Verberkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. 2nd, M. Shah, R. L. Hettich, and J. F. Banfield, “Community proteomics of a natural microbial biofilm.”, *Science (New York, N.Y.)*, vol. 308, no. 5730, pp. 1915-20, Jun. 2005.
[23]N. C. Verberkmoes, A. L. Russell, M. Shah, A. Godzik, M. Rosenquist, J. Halfvarson, M. G. Lefsrud, J. Apajalahti, C. Tysk, R. L. Hettich, and J. K. Jansson, “Shotgun metaproteomics of the human distal gut microbiota.”, *The ISME journal*, vol. 3, no. 2, pp. 179-89, Feb. 2009.
[24]C. A. Kolmeder, M. de Been, J. Nikkilä, I. Ritamo, J. Mättö, L. Valmu, J. Salojärvi, A. Palva, A. Salonen, and W. M. de Vos, “Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions.”, *PloS one*, vol. 7, no. 1, Jan. 2012.
[25]J. D. Rudney, H. Xie, N. L. Rhodus, F. G. Ondrey, and T. J. Griffin, “A metaproteomic analysis of the human salivary microbiota by three-dimensional peptide fractionation and tandem mass spectrometry.”, *Molecular oral microbiology*, vol. 25, no. 1, pp. 38-49, Feb. 2010.
[26]A. R. Erickson, B. L. Cantarel, R. Lamendella, Y. Darzi, E. F. Mongodin, C. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat, J. Raes, N. C. Verberkmoes, C. M. Fraser, R. L. Hettich, and J. K. Jansson, “Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease.”, *PloS one*, vol. 7, no. 11, Jan. 2012.
[27]P. Jagtap, J. Goslinga, J. A. Kooren, T. McGowan, M. S. Wroblewski, S. L. Seymour, and T. J. Griffin, “A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies.”, *Proteomics*, vol. 13, no. 8, pp. 1352-7, Apr. 2013.
[28]J. Kan, T. E. Hanson, J. M. Ginter, K. Wang, and F. Chen, “Metaproteomic analysis of Chesapeake Bay microbial communities.”, *Saline systems*, vol. 1, p. 7, Jan. 2005.
[29]P. Wilmes, M. Wexler, and P. L. Bond, “Metaproteomics provides functional insight into activated sludge wastewater treatment.”, *PloS one*, vol. 3, no. 3, Jan. 2008.
[30]F. Abram, A. Enright, J. O'Reilly, C. H. Botting, G. Collins, and V. O'Flaherty, “A metaproteomic approach gives functional insights into anaerobic digestion.”, *Journal of applied microbiology*, vol. 110, no. 6, pp. 1550-60, Jun. 2011.
[31]A. Siggins, A. Enright, F. Abram, C. Botting, and V. O'Flaherty, “Impact of trichloroethylene exposure on the microbial diversity and protein expression in anaerobic granular biomass at 37°C and 15°C.”, *Archaea (Vancouver, B.C.)*, vol. 2012, p. 940159, Jan. 2012.
[32]A. S. Chuang, Y. O. Jin, L. S. Schmidt, Y. Li, S. Fogel, D. Smoler, and T. E. Mattes, “Proteomic analysis of ethene-enriched groundwater microcosms from a vinyl chloride-contaminated site.”, *Environmental science & technology*, vol. 44, no. 5, pp. 1594-601, Mar. 2010.
[33]R. M. Morris, B. L. Nunn, C. Frazar, D. R. Goodlett, Y. S. Ting, and G. Rocap, “Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction.”, *The ISME journal*, vol. 4, no. 5, pp. 673-85, May 2010.
[34]R. Mayeux, “Biomarkers: potential uses and limitations.”, *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics*, vol. 1, no. 2, pp. 182-8, Apr. 2004.
[35]A. Bjørnstad, B. K. Larsen, A. Skadsheim, M. B. Jones, and O. K. Andersen, “The potential of ecotoxicoproteomics in environmental monitoring: biomarker profiling in mussel plasma using ProteinChip array technology.”, *Journal of toxicology and environmental health. Part A*, vol. 69, no. 1, pp. 77-96, Jan. 2006.
[36]Y. Feng, T. J. Mitchison, A. Bender, D. W. Young, and J. A. Tallarico, “Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds.”, *Nature reviews. Drug discovery*, vol. 8, no. 7, pp. 567-78, Jul. 2009.
[37]M. C. Newman and M. A. Unger, *Fundamentals of Ecotoxicology*. Boca Raton, FL: CRC/Lewis Press.
[38]H. Segner, “Ecotoxicology–How to Assess the Impact of Toxicants in a Multi-Factorial Environment?”, *Multiple stressors: a challenge for the future*, Jan. 2007.

[39]P. Calow and V. Forbes, “Peer Reviewed: does ecotoxicology inform ecological risk assessment?”, *Environmental science & technology*, Jan. 2003.

[40]X. Wang, L. Chang, Z. Sun, Y. Zhang, and L. Yao, “Analysis of earthworm Eisenia fetida proteomes during cadmium exposure: An ecotoxicoproteomics approach.”, *Proteomics*, vol. 10, no. 24, pp. 4476-90, Dec. 2010.

[41]U. Gündel, S. Kalkhof, D. Zitzkat, M. von Bergen, R. Altenburger, and E. Küster, “Concentration-response concept in ecotoxicoproteomics: effects of different phenanthrene concentrations to the zebrafish (Danio rerio) embryo proteome.”, *Ecotoxicology and environmental safety*, vol. 76, no. 2, pp. 11-22, Feb. 2012.

[42]J. Dorts, P. Kestemont, P. Marchand, W. D'Hollander, M. Thézenas, M. Raes, and F. Silvestre, “Ecotoxicoproteomics in gills of the sentinel fish species, Cottus gobio, exposed to perfluorooctane sulfonate (PFOS).”, *Aquatic toxicology (Amsterdam, Netherlands)*, vol. 103, no. 1, pp. 1-8, May 2011.

[43]J. Choi and M. Ha, “Effect of cadmium exposure on the globin protein expression in 4th instar larvae of Chironomus riparius Mg. (Diptera: Chironomidae): an ecotoxicoproteomics approach.”, *Proteomics*, vol. 9, no. 1, pp. 31-9, Jan. 2009.

[44]D. Leroy, E. Haubruge, E. de Pauw, J. P. Thomé, and F. Francis, “Development of ecotoxicoproteomics on the freshwater amphipod Gammarus pulex: identification of PCB biomarkers in glycolysis and glutamate pathways.”, *Ecotoxicology and environmental safety*, vol. 73, no. 3, pp. 343-52, Mar. 2010.

[45]M. Faurobert, E. Pelpoir, and J. Chaïb, “Phenol extraction of proteins for proteomic studies of recalcitrant plant tissues.”, *Methods in molecular biology (Clifton, N.J.)*, vol. 355, pp. 9-14, Jan. 2007.

[46]C. V. Sapan, R. L. Lundblad, and N. C. Price, “Colorimetric protein assay teechniques.”, *Biotechnology and applied biochemistry*, vol. 29, pp. 99-108, Apr. 1999.

[47]T. Zor and Z. Selinger, “Linearization of the Bradford protein assay increases its sensitivity: theoretical and experimental studies.”, *Analytical biochemistry*, vol. 236, no. 2, pp. 302-8, May 1996.

[48]M. M. Bradford, “A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.”, *Analytical biochemistry*, vol. 72, pp. 248-54, May 1976.

[49]K. M. Keiblinger, I. C. Wilhartitz, T. Schneider, B. Roschitzki, E. Schmid, L. Eberl, K. Riedel, and S. Zechmeister-Boltenstern, “Soil metaproteomics - Comparative evaluation of protein extraction protocols.”, *Soil biology & biochemistry*, vol. 54, no. 15, pp. 14-24, Nov. 2012.

[50]H. Hustoft, H. Malerod, S. Wilson, and L. Reubsaet, “A Critical Review of Trypsin Digestion for LC-MS based Proteomics”.

[51]H. Steen and M. Mann, “The ABC‘s (and XYZ’s) of peptide sequencing.”, *Nature reviews. Molecular cell biology*, vol. 5, no. 9, pp. 699-711, Sep. 2004.

[52]B. M. Tissue, (2013, October 7) *CHP - Liquid Chromatography* [Online]. Available: http://www.files.chem.vt.edu/chem-ed/sep/lc/lc.html

[53]I. Miller, J. Crawford, and E. Gianazza, “Protein stains for proteomic applications: which, when, why?”, *Proteomics*, vol. 6, no. 20, pp. 5385-408, Oct. 2006.

[54]E. de Hoffmann and D. Hochstrasser, *Mass Spectrometry,* 3rd ed. Chichester: Wiley.

[55]S. Banerjee and S. Mazumdar, “Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte.”, *International journal of analytical chemistry*, vol. 2012, Jan. 2012.

[56]LamondLabs (2013, October 7) *Electrospray Ionisation (ESI) and Ion Source Overview* [Online]. Available: http://www.lamondlab.com/MSResource/LCMS/MassSpectrometry/electrosprayIonisation.php

[57]S. Nguyen and J. Fenn, “Gas-phase ions of solute species from charged droplets of solutions”, *of the National Academy of Sciences*, Jan. 2007.

[58]R. E. Pedder, “Practical Quadrupole Theory: .Graphical Theory”. Pittsburgh, PA, 28-May-2001.

[59]M. Scigelova, M. Hornshaw, and A. Giannakopulos, “Fourier transform mass spectrometry”, *Molecular & Cellular*, Jan. 2011.

[60]Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. G. Cooks, “The Orbitrap: a new mass spectrometer.”, *Journal of mass spectrometry : JMS*, vol. 40, no. 4, pp. 430-43, Apr. 2005.
[61]R. A. Zubarev, D. M. Horn, E. K. Fridriksson, N. L. Kelleher, N. A. Kruger, M. A. Lewis, B. K. Carpenter, and F. W. McLafferty, “Electron capture dissociation for structural characterization of multiply charged protein cations.”, *Analytical chemistry*, vol. 72, no. 3, pp. 563-73, Feb. 2,”0.
[62]Thermo Fisher Scientific Inc (2013 October 7) *ETD Module* [Online]. Available: http://sjsupport.thermofinnigan.com/TechPubs/manuals/ETD_Start.pdf
[63]Y. Shen, N. Tolić, F. Xie, R. Zhao, S. O. Purvine, A. A. Schepmoes, R. J. Moore, G. A. Anderson, and R. D. Smith, “Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods.”, *Journal of proteome research*, vol. 10, no. 9, pp. 3929-43, Sep. 2011.
[64]Kkmurray (2013 October 7), *Peptide fragmentation notation*, 2nd ed. [Online]. Available: <http://commons.wikimedia.org/wiki/File:Peptide_fragmentation.gif
[65]A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, “Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS ,”d database search.”, *Analytical chemistry*, vol. 74, no. 20, pp. 5383-92, Oct. 2002.
[66]J. S. Cottrell, “Protein identification using MS/MS data.”, *Journal of proteomics*, vol. 74, no. 10, pp. 1842-51, Sep. 2011.
[67]Bin Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, “PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.”, *Rapid communications in mass spectrometry : RCM*, vol. 17, no. 20, pp. 2337-42, Jan. 2003.
[68]J. R. Stevenson and L. L. Bahls, “Periphyton Protocols”, in *Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish*, 2nd ed., no. 6, U.S. Environmental Protection Agency; Office of Water.
[69]S. P. Gygi, G. L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold, “Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 17, pp. 9390-5, Aug. 2000.
[70]B. M. Murphy, S. Swarts, B. M. Mueller, P. van der Geer, M. C. Manning, and M. I. Fitchmun, “Protein instability following transport or storage on dry ice.”, *Nature methods*, vol. 10, no. 4, pp. 278-9, Apr. 2013.
[71]F. Yang, J. Yang, X. Zhang, L. Chen, Y. Jiang, Y. Yan, X. Tang, J. Wang, Z. Xiong, J. Dong, Y. Xue, Y. Zhu, X. Xu, L. Sun, S. Chen, H. Nie, J. Peng, J. Xu, Y. Wang, Z. Yuan, Y. Wen, Z. Yao, Y. Shen, B. Qiang, Y. Hou, J. Yu, and Q. Jin, “Genome dynamics and diversity of Shigella species, the etiologic agents of ,”cillary dysentery.”, *Nucleic acids research*, vol. 33, no. 19, pp. 6445-58, Jan. 2005.
[72]G. Deckert, P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, R. Huber, R. A. Feldman, J. M. Short, G. J. Olsen, and R. V. Swanson, “The complete genome of the hyperthermophilic bacteria Aquifex aeolicus.”, *Nature*, vol. 392, no. 6674, pp. 353-8, Mar. 1998.
[73]L. Sibley, A. Khan, and J. Ajioka, “Genetic diversity of Toxoplasma gondii in animals and humans”, *of the Royal*, Jan. 2009.
[74]R. C. Hightower and R. B. Meagher, “The molecular evolution of actin.”, *Genetics*, vol. 114, no. 1, pp. 315-32, Sep. 1986.
[75]L. Mariño-Ramírez, M. G. Kann, B. A. Shoemaker, and D. Landsman, “Histone structure and nucleosome stability.”, *Expert review of proteomics*, vol. 2, no. 5, pp. 719-29, Oct. 2005.
[76]W. Martin and R. Cerff, “Prokaryotic features of a nucleus-encoded enzyme. cDNA sequences for chloroplast and cytosolic glyceraldehyde-3-phosphate dehydrogenases from mustard (Sinapis alba).”, *European journal of biochemistry / FEBS*, vol. 159, no. 2, pp. 323-31, Sep. 1986.
[77] Thermo Fisher Scientific (2013 October 7), *Eliminate interfering substances from samples for BCA protein assays* [Online]. Available: http://www.piercenet.com/files/TR0008-TCA-acetone-precip.pdf
[78]D. R. Caprette (2013 October 7), *Smeared protein gels* [Online]. Available: http://www.ruf.rice.edu/~bioslabs/studies/sds-page/gelgoofs/smear.html

[79]S. Selvey, E. W. Thompson, K. Matthaei, R. A. Lea, M. G. Irving, and L. R. Griffiths, “Beta-actin--an unsuitable internal control for RT-PCR.”, *Molecular and cellular probes*, vol. 15, no. 5, pp. 307-11, Oct. 2001.

[80]Z. Ning, D. Seebun, B. Hawley, and C. Chiang, “From Cells to Peptides:‘One-Stop’ Integrated Proteomic Processing Using Amphipols”, *Journal of proteome*, Jan. 2013.

12 Appendices

12.1 Appendix 1

This appendix contains a description of how various buffers and solutions were prepared.

All buffers are prepared in milliQ H_2O unless otherwise stated.

Soil extraction buffer
1:1 v:v 50 mM Tris 1% SDS, pH 7.5 + phenol, pH 8.0

Periphyton prep buffer
25 mM tricine
100 mM NaCl
1mM DTT
0.8 ml protease inhibitors (Roche, Complete protease inhibitors)
in 20 ml H_2O

Periphyton solubilisation buffer
7M urea
2M thiourea
2% CHAPS (w/v)
0.5% Triton X-100
30mM IAM
1% (w/v)(64.5 mM) dithiothreitol (DTT)

Periphyton SDT buffer
2 % SDS (w/v), 100 mM DTT in Tris-HCl, pH 7.6

TopTip binding solutions
0.05 % TFA
25 mM KH_2PO_4, 5 % ACN

TopTip releasing solutions
60 % ACN, 0.05 % TFA, prepared fresh
70 % ACN, 0.1 % FA, prepared fresh

FASP solutions
UA: 8 M urea in 0.1 M Tris-HCl pH 8.5
UA-DTT: 100mM DTT in UA
UA-IAM: 50 mM IAM in UA

SDS-PAGE stacking gel (6 %)
9 ml H_2O
3.75 ml stacking buffer (0.5 M Tris-HCl, 0.4 % SDS, pH 6.8)
2.25 ml 40 % AA
12 µl TEMED
120 µl 10 % APS

SDS-PAGE resolving gel (15 %)
5.65 ml H_2O

3.75 ml resolving buffer (1.5 M Tris-HCl, 0.4 % SDS, pH 8.83)
5.65 ml 40 % AA
9 µl TEMED
90 µl 10 % APS

SDS-PAGE running buffer (10x)
0.25 M Tris, 1.9 M Glycin, 1 % SDS

SDS-PAGE sample buffer
4x: 0.5 M SDS
1.8 M sucrose
1.5 mM EDTA
75 mM Tris-HCl
30 % β-mercaptoethanol
bromphenol blue

2x, 1x: the buffer as described above, diluted 2 or 4 times

Coomassie staining solution
40% methanol
50% H_2O
10 % acetic acid
0.5 g Coomassie (G-250, Biorad)
Mixed and filtered to remove colloids.

Destaining solution
40% methanol
50% H_2O
10 % acetic acid

Silver staining, developing solution
6 g Na_2CO_3
200 µl sodium thiosulfate stock (0.13 % w/v)
50 µl formaldehyde 37 %
100 ml milliQ H_2O

Silver staining, pretreatment solution
20 ml sodium thiosulfate stock (0.13 % w/v)
180 ml milliQ H_2O

Silver staining, stop solution
100 ml methanol
76 ml MilliQ H_2O
24 ml acetic acid

Silver staining, silver solution
0.2 g $AgNO_3$
75 µl formaldehyde 37 %
100 ml milliQ H_2O

Silver staining, incubation solution

100 ml milliQ H_2O
100 ml ethanol
2 ml acetic acid

Alkaline-SDS buffer

5 % SDS
50 mM Tris-HCl, pH 8.5
0.15 M NaCl
0.1 mM EDTA
1 mM $MgCl_2$
50 mM DTT

Guanidine buffer

6 M Guanidine HCl, 10 mM DTT in (50 mM Tris, 10 mM $CaCl_2$, pH 7.6)

12.2 Appendix 2

This appendix contains a list of proteins identified from periphyton samples, using various buffers for the extraction.

Proteins identified using prep buffer.

Accession	-10lgP	Coverage (%)	#Peptides (Unique)	Description
Q4YU79|ACT2_PLABA	129.95	8	4(2)	Actin-2 OS=Plasmodium berghei (strain Anka) GN=PB001050.02.0 PE=3 SV=1
P53476|ACT_TOXGO	85.27	9	5(2)	Actin OS=Toxoplasma gondii GN=ACT1 PE=3 SV=1
P10988|ACT1_PLAFO	70.98	12	4(1)	Actin-1 OS=Plasmodium falciparum (isolate NF54) PE=2 SV=1
Q2JTN9|PETD_SYNJA	66.54	5	2(2)	Cytochrome b6-f complex subunit 4 OS=Synechococcus sp. (strain JA-3-3Ab) GN=petD PE=3 SV=2
P15745|GSPF_KLEPN	39.61	4	2(2)	Type II secretion system protein F OS=Klebsiella pneumoniae GN=pulF PE=3 SV=1
Q6GGX3|EBH_STAAR	34.3	1	3(3)	Extracellular matrix-binding protein ebh OS=Staphylococcus aureus (strain MRSA252) GN=ebh PE=4 SV=1
Q2L1K6|SYD_BORA1	33.64	3	1(1)	Aspartate--tRNA ligase OS=Bordetella avium (strain 197N) GN=aspS PE=3 SV=1
A5W992|GLMM_PSEP1	31.64	2	1(1)	Phosphoglucosamine mutase OS=Pseudomonas putida (strain F1 / ATCC 700007) GN=glmM PE=3 SV=1
Q30S46|TGT_SULDN	31.23	2	1(1)	Queuine tRNA-ribosyltransferase OS=Sulfurimonas denitrificans (strain ATCC 33889 / DSM 1251) GN=tgt PE=3 SV=1
A0LR17|CH601_ACIC1	28.34	7	2(2)	60 kDa chaperonin 1 OS=Acidothermus cellulolyticus (strain ATCC 43068 / 11B) GN=groL1 PE=3 SV=1
Q03110|CSP_PLASI	27.68	6	2(2)	Circumsporozoite protein OS=Plasmodium simium GN=CS PE=3 SV=1
P17855|STC2_STAAU	27.21	2	1(1)	Staphylocoagulase OS=Staphylococcus aureus PE=1 SV=1
A4WER4|LSRA_ENT38	27.06	1	1(1)	Autoinducer 2 import ATP-binding protein LsrA OS=Enterobacter sp. (strain 638) GN=lsrA PE=3 SV=1
Q1IWD8|CLPX_DEIGD	26.62	2	1(1)	ATP-dependent Clp protease ATP-binding subunit ClpX OS=Deinococcus geothermalis (strain DSM 11300) GN=clpX PE=3 SV=1
Q5FH94|OBG_EHRRG	26.38	2	1(1)	GTPase obg OS=Ehrlichia ruminantium (strain Gardel) GN=obg PE=3 SV=1
A1SY50|Y2705_PSYIN	26.38	4	1(1)	UPF0229 protein Ping_2705 OS=Psychromonas ingrahamii (strain 37) GN=Ping_2705 PE=3 SV=1
Q06277|IBPA_HAES2	26.27	1	2(2)	Adenosine monophosphate-protein transferase and cysteine protease IbpA OS=Haemophilus somnus (strain 2336) GN=ibpA PE=1 SV=2
Q7U8K7|HISX_SYNPX	26.08	2	1(1)	Histidinol dehydrogenase OS=Synechococcus sp. (strain WH8102) GN=hisD PE=3 SV=1
Q82DM9|RS5_STRAW	25.23	4	1(1)	30S ribosomal protein S5 OS=Streptomyces avermitilis (strain ATCC 31267 / DSM 46492 / JCM 5070 / NCIMB 12804 / NRRL 8165 / MA-4680) GN=rpsE PE=3 SV=1
Q4FM95|TIG_PELUB	25.07	1	1(1)	Trigger factor OS=Pelagibacter ubique (strain HTCC1062) GN=tig PE=3 SV=1
Q7UT69|MOAA_RHOBA	24.94	2	1(1)	Cyclic pyranopterin monophosphate synthase OS=Rhodopirellula baltica (strain SH1) GN=moaA PE=3 SV=1
Q7VDZ2|MURG_PROMA	24.47	2	1(1)	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase OS=Prochlorococcus marinus (strain SARG / CCMP1375 / SS120) GN=murG PE=3 SV=1
C6DET4|SECA_PECCP	24.28	1	1(1)	Protein translocase subunit SecA OS=Pectobacterium carotovorum subsp. carotovorum (strain PC1) GN=secA PE=3 SV=1
P0C6Q3|PGK_VIBCH	23.93	4	1(1)	Phosphoglycerate kinase OS=Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961) GN=pgk PE=3 SV=1
B9M0D5|PROB_GEOSF	23.52	5	1(1)	Glutamate 5-kinase OS=Geobacter sp. (strain FRC-32) GN=proB PE=3 SV=1
Q3Z135|RNB_SHISS	23.45	3	1(1)	Exoribonuclease 2 OS=Shigella sonnei (strain Ss046) GN=rnb PE=3 SV=1
B8E958|SYM_SHEB2	23.11	5	1(1)	Methionine--tRNA ligase OS=Shewanella baltica (strain OS223) GN=metG PE=3 SV=1

A2BUP5|ISPH_PROM5	23.01	2	1(1)	4-hydroxy-3-methylbut-2-enyl diphosphate reductase OS=Prochlorococcus marinus (strain MIT 9515) GN=ispH PE=3 SV=1
A4VUM9|GLYA_STRSY	22.98	2	1(1)	Serine hydroxymethyltransferase OS=Streptococcus suis (strain 05ZYH33) GN=glyA PE=3 SV=1
A0LVY3|GLGC_ACIC1	22.83	3	1(1)	Glucose-1-phosphate adenylyltransferase OS=Acidothermus cellulolyticus (strain ATCC 43068 / 11B) GN=glgC PE=3 SV=1
Q186R0|THIC_CLOD6	22.39	2	1(1)	Phosphomethylpyrimidine synthase OS=Clostridium difficile (strain 630) GN=thiC PE=3 SV=1
P21921|COBL_PSEDE	22.38	2	1(1)	Precorrin-6Y C(5_15)-methyltransferase [decarboxylating] OS=Pseudomonas denitrificans GN=cobL PE=1 SV=1
Q98LF1|COXX1_RHILO	22.04	7	1(1)	Protoheme IX farnesyltransferase 1 OS=Rhizobium loti (strain MAFF303099) GN=ctaB1 PE=3 SV=1
Q98DT6|SSUB1_RHILO	21.55	4	1(1)	Aliphatic sulfonates import ATP-binding protein SsuB 1 OS=Rhizobium loti (strain MAFF303099) GN=ssuB1 PE=3 SV=1
A8HS15|ATPA_AZOC5	21.51	4	1(1)	ATP synthase subunit alpha OS=Azorhizobium caulinodans (strain ATCC 43989 / DSM 5975 / ORS 571) GN=atpA PE=3 SV=1
A1VCX6|HEM3_DESVV	21.39	3	1(1)	Porphobilinogen deaminase OS=Desulfovibrio vulgaris subsp. vulgaris (strain DP4) GN=hemC PE=3 SV=1
Q5WLT3|SYC_BACSK	21.35	3	1(1)	Cysteine--tRNA ligase OS=Bacillus clausii (strain KSM-K16) GN=cysS PE=3 SV=1
B9LKB9|SAT_CHLSY	21.34	2	1(1)	Sulfate adenylyltransferase OS=Chloroflexus aurantiacus (strain ATCC 29364 / DSM 637 / Y-400-fl) GN=sat PE=3 SV=1
Q820I7|PROA_NITEU	21.28	4	1(1)	Gamma-glutamyl phosphate reductase OS=Nitrosomonas europaea (strain ATCC 19718 / NBRC 14298) GN=proA PE=3 SV=1
A3DE52|DXR_CLOTH	21.26	2	1(1)	1-deoxy-D-xylulose 5-phosphate reductoisomerase OS=Clostridium thermocellum (strain ATCC 27405 / DSM 1237) GN=dxr PE=3 SV=1
C6BRT2|PURA_DESAD	21.25	4	1(1)	Adenylosuccinate synthetase OS=Desulfovibrio salexigens (strain ATCC 14822 / DSM 2638 / NCIB 8403 / VKM B-1763) GN=purA PE=3 SV=1
B3E4H8|LPXB_GEOLS	21.22	7	1(1)	Lipid-A-disaccharide synthase OS=Geobacter lovleyi (strain ATCC BAA-1151 / DSM 17278 / SZ) GN=lpxB PE=3 SV=1
Q9KUC0|PBPB_VIBCH	21.04	1	1(1)	Penicillin-binding protein 1B OS=Vibrio cholerae serotype O1 (strain ATCC 39315 / El Tor Inaba N16961) GN=mrcB PE=3 SV=1
Q7DDB6|Y1497_NEIMB	20.92	1	1(1)	Probable TonB-dependent receptor NMB1497 OS=Neisseria meningitidis serogroup B (strain MC58) GN=NMB1497 PE=1 SV=1
Q8CNW8|DHA_STAES	20.78	4	1(1)	Alanine dehydrogenase OS=Staphylococcus epidermidis (strain ATCC 12228) GN=ald PE=3 SV=1
A2SBW1|PYRE_METPP	20.71	3	1(1)	Orotate phosphoribosyltransferase OS=Methylibium petroleiphilum (strain PM1) GN=pyrE PE=3 SV=1
Q48FC2|Y3775_PSE14	20.67	4	1(1)	Probable transcriptional regulatory protein PSPPH_3775 OS=Pseudomonas syringae pv. phaseolicola (strain 1448A / Race 6) GN=PSPPH_3775 PE=3 SV=1
Q8EUA6|FTSH_MYCPE	20.56	3	1(1)	ATP-dependent zinc metalloprotease FtsH OS=Mycoplasma penetrans (strain HF-2) GN=ftsH PE=3 SV=1
B1ZZ32|MNMA_OPITP	20.43	3	1(1)	tRNA-specific 2-thiouridylase MnmA OS=Opitutus terrae (strain DSM 11246 / PB90-1) GN=mnmA PE=3 SV=1
B0TM18|RPOC_SHEHH	20.12	2	1(1)	DNA-directed RNA polymerase subunit beta' OS=Shewanella halifaxensis (strain HAW-EB4) GN=rpoC PE=3 SV=1
A3NYC9|URE1_BURP0	20.07	2	1(1)	Urease subunit alpha OS=Burkholderia pseudomallei (strain 1106a) GN=ureC PE=3 SV=1
A7Z0D4|PDXT_BACA2	20.04	7	1(1)	Glutamine amidotransferase subunit PdxT OS=Bacillus amyloliquefaciens (strain FZB42) GN=pdxT PE=3 SV=1

Proteins identified using the solubilisation buffer.

Accession	**-10lgP**	**Coverage (%)**	**#Peptides (Unique)**	**Description**
P05098|PHEA_FREDI	167.01	38	5(2)	C-phycoerythrin alpha chain OS=Fremyella diplosiphon GN=cpeA PE=1 SV=1
P53476|ACT_TOXGO	151.15	14	3(1)	Actin OS=Toxoplasma gondii GN=ACT1 PE=3 SV=1
P05097|PHEB_FREDI	138.06	27	4(3)	C-phycoerythrin beta chain OS=Fremyella diplosiphon GN=cpeB PE=1 SV=1
P13530|PHCA_SYNE7	124.86	17	2(1)	C-phycocyanin alpha chain OS=Synechococcus elongatus (strain PCC 7942) GN=cpcA1 PE=1 SV=2
Q7RME1|ACT1_PLAYO	115.51	14	3(1)	Actin-1 OS=Plasmodium yoelii yoelii GN=PY02240 PE=3 SV=1

P26183|ACT_CRYPV	109.33	11	3(1)	Actin OS=Cryptosporidium parvum PE=3 SV=1
Q4FP38|ATPB_PELUB	97.32	8	3(1)	ATP synthase subunit beta OS=Pelagibacter ubique (strain HTCC1062) GN=atpD PE=3 SV=1
Q4YU79|ACT2_PLABA	90.26	6	1(1)	Actin-2 OS=Plasmodium berghei (strain Anka) GN=PB001050.02.0 PE=3 SV=1
P53469|ACT2_OXYTR	90.24	7	2(2)	Actin_ cytoplasmic OS=Oxytricha trifallax PE=3 SV=1
Q02179|PHEA1_SYNPY	84.92	20	2(2)	C-phycoerythrin class 1 subunit alpha OS=Synechococcus sp. (strain WH8020) GN=cpeA PE=1 SV=1
P07325|PHAA_ANACY	77.77	6	1(1)	Allophycocyanin alpha chain OS=Anabaena cylindrica GN=apcA PE=1 SV=1
A2BTI9|PSAC_PROMS	77.1	32	2(1)	Photosystem I iron-sulfur center OS=Prochlorococcus marinus (strain AS9601) GN=psaC PE=3 SV=1
Q54715|PHCA_SYNY3	67.87	12	1(1)	C-phycocyanin alpha chain OS=Synechocystis sp. (strain PCC 6803 / Kazusa) GN=cpcA PE=1 SV=1
Q10Z38|ATPB_TRIEI	67.67	4	2(1)	ATP synthase subunit beta OS=Trichodesmium erythraeum (strain IMS101) GN=atpD PE=3 SV=1
Q39KX8|ATPA_BURS3	67.38	2	1(1)	ATP synthase subunit alpha OS=Burkholderia sp. (strain 383) GN=atpA PE=3 SV=1
Q5P7G2|CH60_AROAE	63.12	5	2(1)	60 kDa chaperonin OS=Aromatoleum aromaticum (strain EbN1) GN=groL PE=3 SV=1
B2J8T2|RBL_NOSP7	61.31	2	1(1)	Ribulose bisphosphate carboxylase large chain OS=Nostoc punctiforme (strain ATCC 29133 / PCC 73102) GN=cbbL PE=3 SV=1
Q2JTQ2|RL7_SYNJA	58.76	9	1(1)	50S ribosomal protein L7/L12 OS=Synechococcus sp. (strain JA-3-3Ab) GN=rplL PE=3 SV=1
B3EJK9|ATPB_CHLPB	47.6	4	1(1)	ATP synthase subunit beta OS=Chlorobium phaeobacteroides (strain BS1) GN=atpD PE=3 SV=1
Q747A2|SYC_GEOSL	43.65	4	1(1)	Cysteine--tRNA ligase OS=Geobacter sulfurreducens (strain ATCC 51573 / DSM 12127 / PCA) GN=cysS PE=3 SV=1
A6GYU1|RL7_FLAPJ	41.38	10	1(1)	50S ribosomal protein L7/L12 OS=Flavobacterium psychrophilum (strain JIP02/86 / ATCC 49511) GN=rplL PE=3 SV=1
A3M1Z8|PROA_ACIBT	41.16	3	1(1)	Gamma-glutamyl phosphate reductase OS=Acinetobacter baumannii (strain ATCC 17978 / NCDC KC 755) GN=proA PE=3 SV=2
A5GWJ9|RL7_SYNR3	40.7	9	1(1)	50S ribosomal protein L7/L12 OS=Synechococcus sp. (strain RCC307) GN=rplL PE=3 SV=1
A8LJR6|ATPA2_DINSH	37.17	2	1(1)	ATP synthase subunit alpha 2 OS=Dinoroseobacter shibae (strain DFL 12) GN=atpA2 PE=3 SV=1
Q10701|HELY_MYCTU	35.06	1	1(1)	Probable helicase HelY OS=Mycobacterium tuberculosis GN=helY PE=3 SV=1
P23324|UBIQP_EUPEU	34.3	4	1(1)	Polyubiquitin OS=Euplotes eurystomus PE=3 SV=2
B4TET8|MDTH_SALHS	33.66	3	1(1)	Multidrug resistance protein MdtH OS=Salmonella heidelberg (strain SL476) GN=mdtH PE=3 SV=1
Q28VY6|MUTS_JANSC	33.36	1	1(1)	DNA mismatch repair protein MutS OS=Jannaschia sp. (strain CCS1) GN=mutS PE=3 SV=1

Proteins identified using solubilisation + acetic acid extraction.

Accession	-10lgP	Coverage (%)	#Peptides (Unique)	Description
P26183|ACT_CRYPV	75.73	3	1(1)	Actin OS=Cryptosporidium parvum PE=3 SV=1
B0B9Z5|SYD_CHLTB	47.39	2	1(1)	Aspartate--tRNA ligase OS=Chlamydia trachomatis serovar L2b (strain UCH-1/proctitis) GN=aspS PE=3 SV=1
A1WBI7|PANB_ACISJ	47.33	2	1(1)	3-methyl-2-oxobutanoate hydroxymethyltransferase OS=Acidovorax sp. (strain JS42) GN=panB PE=3 SV=1

Proteins identified using SDT buffer.

Accession	-10lgP	Coverage (%)	#Peptides (Unique)	Description
P08040|PHCA2_FREDI	69.25	8	1(1)	C-phycocyanin-2 alpha chain OS=Fremyella diplosiphon GN=cpcA2 PE=2 SV=1
Q2K3G8|ATPA_RHIEC	65.63	2	1(1)	ATP synthase subunit alpha OS=Rhizobium etli (strain CFN 42 / ATCC 51251) GN=atpA PE=3 SV=1
A2C4I4|ATPB_PROM1	44.95	3	1(1)	ATP synthase subunit beta OS=Prochlorococcus marinus (strain NATL1A) GN=atpD PE=3 SV=1