Metaproteogenomics-guided enzyme discovery
Targeted identification of novel proteases in microbial communities

Mikaela Johansson

Department of Physics, Chemistry and Biology
Linköping University, Sweden

Linköping 2018
Cover: 2-D DIGE gel used for the identification of extracellular proteases secreted by members of a methanogenic microbial community.
If we knew what it was we were doing, it would not be called research, would it.

Albert Einstein
Abstract

Industrial biotechnology is a large and growing industry as it is part of establishing a “greener” and more sustainable bioeconomy-based society. Using enzymes as biocatalysts is a viable alternative to chemicals and energy intense industrial processes and is en route to a more sustainable industry. Enzymes have been used in different areas for ages and are today used in many industrial processes such as biofuels production, food industry, tanning, chemical synthesis, pharmaceuticals etc. Enzymes are today a billion-dollar industry in itself and the demand for novel catalysts for various present and future processes of renewable resources are high and perfectly in line with converting to a more sustainable society.

Most enzymes used in industry today have been identified from isolated and pure cultured microorganisms with identified desirable traits and enzymatic capacities. However, it is known that less than 1% of all microorganisms can be obtained in pure cultures. Thus, if we were to rely solely on pure culturing, this would leave the 99% of the microorganisms that constitute the “microbial dark matter” uninvestigated for their potential in coding for and producing valuable novel enzymes. Therefore, to investigate these “unculturable” microorganisms for novel and valuable enzymes, pure-culture independent methods are needed.

During the last two decades there has been a fast and extensive development in techniques and methods applicable for this purpose. Especially important has been the advancements made in mass spectrometry for protein identification and next generation sequencing of DNA. With these technical developments new research fields of proteomics and genomics have been developed, by which the complete protein complement of cells (the proteome) and all genes (the genome) of organisms can be investigated. When these techniques are applied to microbial communities these fields of research are known as meta-proteomics and meta-genomics.

However, when applied to complex microbial communities, difficulties different from those encountered in their original usage for analysis of single multicellular organisms or cell linages arises, and when used independently both methods have their own limitations and bottlenecks. In addition, both metaproteomics and metagenomics are largely non-targeting techniques. Thus, if the purpose is still to - somewhat contradictory – use these non-targeting methods for targeted identification of novel enzymes with
certain desired activities and properties from within microbial communities, special measures need to be taken.

The work presented in this thesis describes the development of a method that combines metaproteomics and metagenomics (i.e. metaproteogenomics) for the targeted discovery of novel enzymes with desired activities, and their correct coding genes, from within microbial communities. Thus, what is described is a method that can be used to circumvent the pure-culturing problem so that a much larger fraction of the microbial dark matter can be specifically investigated for the identification of novel valuable enzymes.
Populärvetenskaplig sammanfattning

Industriell bioteknik är en stor och växande industri eftersom den är en del av att etablera ett ”grönare” och mer hållbart och bioekonomibaserat samhälle. Att använda enzymer som biokatalysatorer är ett mycket användbart alternativ till kemikalier och energintensiva industriella processer på vägen mot en mer hållbar industri. Enzymer har använts i alla tider i olika processer och används i stor skala idag i många industrier som t.ex. biobränsleproduktion, livsmedelsindustrin, lädergarvning, kemisk syntes, läkemedel etc. Enzymer är i sig själva en miljardindustri och efterfrågan på nya enzymer för olika nuvarande och framtida processer för behandling av förnyelsebara råvaror är stor, och faller väl i linje med övergången till ett mer hållbart samhälle.

De flesta enzymer som används inom industri idag har identifierats genom isolering och renodling av mikroorganismer som har identifierats ha önskvärda egenskaper och enzymatisk kapacitet. Det är dock känt att mindre än 1% av alla mikroorganismer går att renodla. Detta betyder alltså att om vi enbart ska förlita oss på renodling så kommer 99% av alla mikroorganismer som utgör den ”mikrobiella mörka materian” inte kunna undersökas för sin potential att koda för och producera värdefulla nya enzymer. För att kunna undersöka dessa ”icke odlingsbara” mikroorganismerna och få fram nya och värdefulla enzymer behövs därför metoder som är renodlingsberoende.

De senaste två årtiondena har det varit en snabb och omfattande utveckling av tekniker och metoder som är tillämpbara för just detta syfte. Av särskild vikt har framsteg inom masspektrometri för proteinidentifiering och massiv parallell DNA sekvensering (next generation sequencing) varit. Denna tekniska utvecklingen har gått hand i hand med framväxten av nya forskningsfält som proteomik och genomik, där man studerar alla proteiner (proteomet) hos celler och alla gener (genomet) hos organismer. När dessa tekniker appliceras på mikrobiella samhällen kallas forskningsfälten för meta-proteomik och meta-genomik.

Dock, när dessa tekniker appliceras på komplexa mikrobiella samhällen uppstår andra svårigheter än när de används för sina ursprungliga syften att analysera enskilda flercelliga organismer eller cellinjer, och när de används på egen hand har båda teknikerna sina egna begränsningar och flaskhalsar. Ovanpå detta är både metaproteomik och metagenomik till stor del icke-riktade metoder. Om målet ändå är att - något motsägelsefullt - använda dessa icke-riktade metoder för att riktat identifiera nya
Enzymer med vissa önskade aktiviteter och egenskaper ur komplexa mikrobiella samhällen, behövs därför särskilda åtgärder tas.

Arbetet som presenteras i denna avhandling beskriver utvecklingen av en metod som kombinerar metaproteomik och metagenomik (d.v.s. metaproteogenomik) för den riktade identifieringen av nya enzymer med önskade aktiviteter, och deras korrepta kodande gener, ur mikrobiella samhällen. Därmed beskrivs en metod som kan användas för att kringgå renodlingsproblemet så att en mycket större del av den mikrobiella mörka materian kan specifikt undersökas för identifiering av nya värdefulla enzymer.
List of publications included

This thesis is based on the following papers, which are referred to in the text by their Roman numerals (I-IV).


 (* These authors contributed equally to the work)


 (* These authors contributed equally to the work)

III Karlsson, M., Johansson, M.A. and Speda, J. 'Shotgun metagenomic analysis of community structure of co-cultured microbial communities under controlled artificial conditions: Implications and applications.'

 *In manuscript.*


 *In manuscript.*
Contribution report

I  Mikaela Johansson (MJ) participated in all planning, designed and executed the protease induction experiment and co-authored the manuscript.

II  MJ was involved in all planning and executed all protein extraction experiments. MJ further contributed to the interpretation of the results and co-authored the manuscript.

III  MJ was involved in the analysis and interpretation of metagenomic data and contributed to the final editing of the manuscript.

IV  MJ was involved in all planning and performed all the experiments. MJ also did the majority of data analysis, interpretation of results and wrote the majority of the manuscript.
Conference contributions


Paper not included in the thesis

Thesis committee

Graduate supervisor
Dr. Martin Karlsson
Department of Physics, Chemistry and Biology
Division of Molecular Biotechnology
Linköping University, Linköping, Sweden

Co-supervisor
Professor emeritus Bengt-Harald (Nalle) Jonsson
Department of Physics, Chemistry and Biology
Division of Molecular Biotechnology
Linköping University, Linköping, Sweden

Faculty opponent
Dr. Florian-Alexander Herbst
Eurofins Dr. Specht Laboratorien GmbH, Hamburg, Germany

Committee board
Professor Libeth Olsson
Department of Chemical and Biological engineering
Division of Biotechnology
Chalmers, Gothenburg, Sweden

Dr. Karin Enanader
Department of Physics, Chemistry and Biology
Division of Molecular Physics
Linköping University, Linköping, Sweden

Dr. Alex Enrich Prast
TEMA- Department of Thematic Studies (TEMA)
Environmental Change (TEMAM)
Linköping University, Linköping, Sweden
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D</td>
<td>one-dimensional</td>
</tr>
<tr>
<td>1-DE</td>
<td>one-dimensional gel electrophoresis</td>
</tr>
<tr>
<td>2-D</td>
<td>two-dimensional</td>
</tr>
<tr>
<td>2-DE</td>
<td>two-dimensional gel electrophoresis</td>
</tr>
<tr>
<td>2-D DIGE</td>
<td>two-dimensional difference gel electrophoresis</td>
</tr>
<tr>
<td>API</td>
<td>active pharmaceutical ingredient</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>CCD</td>
<td>charge-coupled device</td>
</tr>
<tr>
<td>CHAPS</td>
<td>3-[[3-cholamidopropyl]dimethylammonio]-1-propanesulfonate</td>
</tr>
<tr>
<td>CID</td>
<td>collision induced dissociation</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DIGE</td>
<td>difference gel electrophoresis</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiotreitol</td>
</tr>
<tr>
<td>ESI</td>
<td>electrospray ionization</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>IEF</td>
<td>isoelectric focusing</td>
</tr>
<tr>
<td>IPG</td>
<td>immobilized pH gradient</td>
</tr>
<tr>
<td>kat</td>
<td>katal</td>
</tr>
<tr>
<td>LC</td>
<td>liquid chromatography</td>
</tr>
<tr>
<td>LCFA</td>
<td>long chain fatty acids</td>
</tr>
<tr>
<td>LIT</td>
<td>linear ion trap</td>
</tr>
<tr>
<td>LTQ</td>
<td>linear trap quadrupole</td>
</tr>
<tr>
<td>MALDI</td>
<td>matrix assisted laser desorption/ionization</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MS/MS</td>
<td>tandem mass spectrometry</td>
</tr>
<tr>
<td>MW</td>
<td>molecular weight</td>
</tr>
<tr>
<td>m/z</td>
<td>mass/charge ratio</td>
</tr>
<tr>
<td>NGS</td>
<td>next-generation sequencing</td>
</tr>
<tr>
<td>NHS</td>
<td>N-hydroxylsuccinimidylerase</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>OLR</td>
<td>organic loading rate</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PEG</td>
<td>polyethylene glycol</td>
</tr>
<tr>
<td>pI</td>
<td>isoelectric point</td>
</tr>
<tr>
<td>PTM</td>
<td>post-translational modification</td>
</tr>
<tr>
<td>Q</td>
<td>quadrupole</td>
</tr>
<tr>
<td>QIT</td>
<td>quadrupole time of flight</td>
</tr>
<tr>
<td>SDS PAGE</td>
<td>sodium dodecyl sulfate polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>TCA</td>
<td>trichloroacetic acid</td>
</tr>
<tr>
<td>TOF</td>
<td>time of flight</td>
</tr>
<tr>
<td>U</td>
<td>enzyme activity unit</td>
</tr>
<tr>
<td>VFA</td>
<td>volatile fatty acids</td>
</tr>
</tbody>
</table>
Table of Contents

Abstract ........................................................................................................ V
Populärvetenskaplig sammanfattning ....................................................... VII
List of publications included................................................................... IX
Contribution report ................................................................................... X
Conference contributions ....................................................................... XI
Paper not included in the thesis ........................................................... XII
Thesis committee .................................................................................... XIII
Abbreviations ........................................................................................ XV
Preface ..................................................................................................... XIX
Introduction ........................................................................................... 1
  Biotechnology to industrial biotechnology ........................................ 1
  Industrial biotechnology ....................................................................... 2
  Enzymes in industrial biotechnology .................................................. 2
  Enzyme structure ................................................................................ 4
  Hydrolytic enzymes ............................................................................. 4
    Amylases ............................................................................................ 4
    Cellulases .......................................................................................... 5
  Lipases ................................................................................................. 5
  Proteases ............................................................................................. 5
  The importance of proteases in industry ............................................ 9
Enzyme discovery .................................................................................... 9
  Metagenomics ..................................................................................... 11
  Metaproteomics .................................................................................. 14
    Combining meta-omics techniques into metaproteogenomics ........ 16
Biogas producing microbial community ............................................. 17
  Metaproteogenomics for discovery of novel enzymes ...................... 19
Aims of the study .................................................................................. 25
Methods ............................................................................................... 27
Preface

It has been seven years since I was introduced to the world of enzyme discovery during my master thesis work. It was still the beginning of the development of a new approach that could be used for enzyme discovery. This work has been challenging since there was no other group doing the same thing but that has also been the greatest reward when the solutions have come.

The work presented in this thesis covers the research done during my graduate studies. It describes the development of an approach using anaerobic microbial communities for discovering new enzymes, mainly proteases. It gives an introduction into the subject giving the background for the project and putting it into perspective. All methods used are described and the articles further detailing the work is included at the end of the thesis.

Being part of this journey has been amazing and has given me the opportunity to be part of developing an approach which I feel will be very useful and contribute to the field of enzyme discovery.

I hope you enjoy reading

Mikaela Johansson
Linköping, April 2018
Don’t try this at home.
Introduction

Biotechnology to industrial biotechnology

Biotechnology is “the application of scientific and engineering principles to the processing of materials by biological agents”[4]. Fermentation, a type of biotechnology, has been used for centuries to produce different products such as beer, bread, wine and pickles. Originally the processes involving microbes and fungi were used for preservation of food such as, vegetables, fruit and milk but then developed into other products, such as cheese and spirits, which was for pleasure and satisfaction. The concept of catalysis and catalysts was created by the Swedish chemist Jacob Berzelius in 1835 [5]. With the concept of catalysis came huge progress in the investigations of reaction rates and catalytic processes. Catalysts became sought after to accelerate reaction rates and many industries in the 19’th and 20’th centuries are built around the concept of catalytic processes [6].The first world war brought a huge wave of development for biotechnology. The acetone-butanol fermentation was developed in England by Weizmann, and the glycerol fermentation was developed in Germany by Neuberg. Both these products were very important in the production of making munitions to support the war effort on both sides. Following WWI biotechnology made great progress with
important developments in fermentation, bioconversions and enzymatic processes, with the most well-known and probably most important being the discovery of penicillin by Fleming and its development by Florey, Heatley, Chain and Abraham. In the 1970’s another large wave came in biotechnology accompanied by the first oil crisis, bringing light to the need for renewable materials. One of the most important technologies to emerge during this period was the introduction of recombinant DNA technology by Berg, Byer and Cohen in 1972 [7]. The modern industrial biotechnology came from the merger of molecular biology and industrial microbiology [8]. Today enzymes are used to catalyze reactions in many different industries, such as food, leather, detergent, paper and pulp, agriculture, pharmaceuticals, textile and organic synthesis to mention some [9-12].

**Industrial biotechnology**

Currently, our society is facing a number of environmental challenges and have to develop new ways to produce fuel and energy in order to reduce our dependence on fossil fuels [13]. This also includes making industries “greener” in order to reduce pressure on the environment. Industrial biotechnology is already a large part of this transition and will most likely be the leading sector in bringing a more sustainable industry [4].

Industrial biotechnology was announced as one of six key enabling technologies by the European commission in 2009 because of its possibility to “bring cleaner and sustainable process alternatives for industrial and agri-food operations” and also to facilitate the replacement of using non-renewable materials with renewables in industry [14].

Industrial biotechnology encompasses a variety of applications and tools such a traditional “bioprocesses” and production of bio-based commodities like plastics, chemicals and fuel. Industrial biotechnology also encompasses the microorganisms and enzymes, both natural and genetically engineered as well as the processes in which they are used. These areas are a large and growing part of industrial biotechnology and there are many companies and research groups devoted to developing them for commercial use [15]. Enzymes have quickly gained interest due to their applicability in a large variety of industries. The global market for industrial enzymes in 2014 was estimated to $4.2 billion and expected to reach $6.3 billion in 2020 [16].

**Enzymes in industrial biotechnology**

Processes mediated by enzymes are of interest due to their ability to reduce process time, lower energy consumption, increase cost effectiveness, while at the same time being non-
toxic and environmentally friendly. Additionally, enzymes and microbes can, thanks to recombinant DNA technology, be engineered so that they become, for example, more stable or produce at higher rates [16, 17]. A popular method to overcome the limitations of rational design for procuring new stable and active enzyme variants is to use directed evolution where improvements of biocatalysts are made, in the lab, by mimicking and speeding up the Darwinian evolution [18, 19]. The principle of directed evolution starts with a gene, or genes, of interest coding for a specific protein or proteins which is randomly mutated to produce a large pool of protein variants. These variants are then expressed, subsequently screened and selected in order to isolate the protein with the sought after characteristics, if this does not occur another round of mutations are performed until the desired change has been introduced [18].

According to Jemli et. al [20], industrial enzymes can be divided into three groups: technical, food and animal feed. Technical enzymes are used in applications in various industries such as; detergent, textile, paper, fuel and alcohol. It is to this segment that the largest group of commercialized enzymes belong. The second largest group is food enzymes and includes industries such as; dairy, brewing, wine, juice, fat and oil and baking. The third group is the animal feed industry. The enzymes used in industry are mostly hydrolytic enzymes with proteases being the major enzyme type followed by carbohydrases and lipases [21]. Most enzymes in industry today are derived from mesophilic microbes and fungi and therefore operate best under mild conditions, close to neutral pH, normal pressure and in aqueous solutions [22]. This means that the industrial processes using these enzymes operate under very mild conditions which is part of what makes these processes environmentally friendly. However, these conditions are not always applicable, some industrial processes need to operate under high pressure, high temperatures or non-aqueous solvents etc. For such industries to become environmentally friendly there is a need for enzymes that can operate under harsh conditions [23]. The introduction of genetic modifications in a microbe or an enzyme which normally functions in mild conditions, so that it will be highly active and stable in extreme conditions is at present a daunting task [24]. An alternative to genetic modifications is to search for novel enzymes in extremophilic microorganisms, i.e. microbes evolved to survive under extreme conditions. They include a number of different classes such as; thermophiles, acidophiles, alkalophiles, psychrophiles, halophiles and barophiles and are able to thrive in various ecological niches such as deep-sea hydrothermal vents, hot springs and sulfataric fields [22]. The extremophiles are naturally evolved to survive harsh conditions and already harbor enzymes with properties that are sought after by industries. Enzymes that have the largest share of the industrial
enzyme market, and are of most interest, are hydrolytic enzymes such as proteases, lipases, amylases and other carbohydrates which together constitute 75% of all industrial enzymes [25, 26].

Thus, enzymes are used in a large variety of industries and enzymes that are more efficient, stable or able to catalyze new reactions are highly sought after.

**Enzyme structure**

Enzymes are proteins that are able to catalyze a reaction by lowering the activation energy of a reaction without being consumed in the process. Proteins are made up of amino acids forming a polypeptide chain that folds into a three-dimensional structure, forming a mature enzyme. Protein structures are described on four levels, the polypeptide chain make up the primary structure. The secondary structure describes the local structure, such as α-helices and β-strands. The folding of the polypeptide chain driven by interactions between different secondary structure elements into the proteins native structure, is called the tertiary structure. The quaternary structure describes the interactions or aggregations of separate polypeptide chains [27, 28]. Enzymes have a catalytic site that recognizes specific sequences or motifs of substrates and are able to specifically bind the substrates and the reaction intermediates but have a low affinity for the product. This could be anything from isomerization or hydrolysis to ligation and oxidation/reduction [29]. Without enzymes life would not exist since most reactions take too long to occur spontaneously under standard conditions.

**Hydrolytic enzymes**

Hydrolytic enzymes break polymers into smaller fractions by facilitating hydrolysis which is the reaction where a chemical bond is broken by the addition of water. In proteases, for instance, this means to break the peptide bond. There are a variety of hydrolytic enzymes such as; cellulases, chitinases, lipases, amylases, pectinases and proteases. Since the focus of this thesis is on proteases they will be described more in depth. Some of the most common industrial enzyme classes are described briefly below.

**Amylases**

The most commonly used enzymes in industry after proteases are amylases. Amylases, in particular α-amylases, are starch-degrading enzymes that catalyze the hydrolysis of the internal α-1,4-O-glycosidic bond in polysaccharides retaining the α-anomeric configuration in the products [30]. Amylases, having approximately 25% of the world
enzyme market [31] are of great importance in biotechnology because of its wide range of applications in industries such as biofuels, food, fermentation, textiles and paper industry [32]. Amylases can be obtained from a variety of sources such as plants, animals and microbes. Microbial amylases are the ones preferred because they are easiest to produce in large quantities. Amylases have in practice replaced the chemical hydrolysis of starch in the starch industry [33].

**Cellulases**

Cellulases are members of the glycoside hydrolase family and are responsible for hydrolyzing the $\beta$-1,4 glycosidic bonds in cellulose [34]. Cellulose is the most abundant polymer in plants. Being able to convert cellulose biomass by using cellulases would unlock a large renewable source of glucose to be used for production of fuels, chemicals and food and feed [35]. Cellulases are thus of great interest for industry and biotechnology companies [36].

**Lipases**

Lipases is a class of hydrolases able to hydrolyze triglycerides to fatty acid and glycerol at the oil-water interface. Some lipases are also able to catalyze enantioselective hydrolysis reactions and transesterification [37, 38]. Lipases are of great interest for industrial purposes due to their numerous application in industries such as processing of fats and food, for use in detergents and in the synthesis of chemicals [39]. Lipases are considered the third largest industrial enzyme group after proteases and carboxydrases [40] and have become the mostly used group of enzymatic catalyst in organic chemistry [41].

**Proteases**

Proteases are a group of enzymes known as peptidyl-peptide hydrolases and catalyze the cleavage of the peptide bond in proteins using proteolysis [42, 43]. All proteins undergo proteolytic modification of some kind during their lifetime, it can be for maturation, during synthesis or for degradation [44]. This makes proteases vital for all processes affected by proteins and for life to exist. It also means that proteases are very versatile in their function and thus the architectural design of a protease varies from single units protein of $\sim 20 \text{kDa}$ to large proteasomes of 0.7-6 MDa [45]. Proteases make up the largest family of enzymes and constitute 2% of the human genome and most other genomes [46]. Proteases are part of all forms of life and can therefore be derived from any organism be it animals, plants or microorganisms. However, due to the huge demand of enzymes needed for industrial use, microorganisms and especially bacteria, have become the main
source for protease production [47, 48]. This thesis will focus on microbial proteases, since they are of greatest interest for the industry.

Protease classification

Proteases are classified according to their activity and are numbered according to enzyme commission system (EC) which was established in 1955. Proteases are classified under hydrolases which is group EC 3 and then further under the subfamily EC 3.4 which is hydrolases which act on peptide bonds, peptidases. Proteases encompass EC 3.4.1-25 and 99. However, many proteases are still referred to with their non-systematic names, such as trypsin, papain, renin or alcalase [49]. The systematic name is peptide hydrolases, but the enzymes are commonly called, proteases, proteolytic enzymes, proteinases and peptidases, in this thesis they will be referred to as proteases. Proteases are also commonly divided into four subclasses, which are not connected to the EC system but is based on evolutionarily relationships. The subclasses, are serine-, cysteine-, aspartic- and metallo-proteases and comes from the critical amino acid involved in catalysis, with metalloproteases having a metal ion as cofactor [49, 50]. Depending on where the protease cleaves a substrate they are named as either exo- or endopeptidases. Exopeptidases act near the ends of the polypeptide chain and are thus named amino- or carboxypeptidases depending on which end they act upon and release single amino acids, dipeptides or tripeptides. Endopeptidases act further in on the polypeptide chain [51]. Proteins have also been classified by homology of the “peptidase unit” where the active site residues are located in the MEROPS database which was established in 1993 (https://www.ebi.ac.uk/merops/) where the proteases are grouped into families based on similarity and further clustered into clans depending on tertiary structure homology [52]. There are currently six families on MEROPS classified according to their catalytic type, aspartic, cysteine, glutamic, metallo, serine and threonine, and also a category for proteases of still unknown catalytic function.
Protease specificity

Proteases have narrow or broad specificity in their active sites and are in many cases able to cleave different substrates, and a single substrate can be cleaved by a multitude of proteases [44, 53]. Proteases can be very promiscuous or very selective in their specificity against substrates. Proteases does not recognize a specific substrate but binding sequences and specificity is determined by the molecular interactions at the interface of protease-substrate binding [53]. A nomenclature based on cleavage-site specificity has been developed by Schechter and Berger [2]. The amino-acid residues in the protease where binding occur are called subsites (S) and are numbered from the active site and outwards, S1 N-terminal and S1’ C-terminal, the residues that are recognized in the substrate are called peptides (P) and are numbered P1 and P1’ from the cleavage site and outwards. Adapted from [2].

Figure 1. Protease specificity model. The amino acid residues in the protease where binding occur are called subsites (S) and are numbered from the active site and outwards, S1 N-terminal and S1’ C-terminal, the residues that are recognized in the substrate are called peptides (P) and are numbered P1 and P1’ from the cleavage site and outwards. Adapted from [2].
Protease mechanisms

Hydrolysis of peptides has been described as an acid-base reaction and in some cases a nucleophilic attack on the peptide bond of the substrate leads to an intermediary complex called “acyl intermediate”. The complex breaks down rapidly and transfers a proton to a residue in the protease acting as a general base allowing for a water molecule to hydrolyze the peptide bond. It is the nature of the initial nucleophilic attack which differs between different catalytic types of proteases. The catalytic types can be divided into those that use an amino-acid residue for the nucleophilic attack (protein nucleophiles) these are serine and threonine proteases where the hydroxyl group act as a nucleophile,
or the cysteine proteases where the thiol group is the nucleophile. Those that use an activated water molecule for the nucleophilic attack (water nucleophiles), where the water molecule can either be activated by the aspartic or glutamic amino-acid residues or by a metallic ion bound as a co-factor constituting the aspartyl-, glutamyl- and metalloproteases (figure 2) [3, 54, 55].

The importance of proteases in industry

Proteases are of significant importance to the industry and encompasses 60% of the global enzyme market [56, 57]. As stated above the microbial proteases are considered to have the highest value since their production capacity is able to meet global demand as well as the microbes being a renewable source in themselves [58, 59]. Proteases are a major part of the detergent industry [60, 61] as well as the food, dairy, leather and pharmaceutical industries [62]. A growing trend is the use of microorganisms for waste management in order to convert wastes into biomass, and new proteases will be needed for this application [63]. Most bacterial proteases used in industry today comes from the species *Bacillus* with 50% of the total enzyme market. The alkaline serine proteases, subtilisins, is the most dominant commercial enzyme in itself due to its use in detergents [64-66]. With more industries using proteases in their processes and products due to the benefits of using enzymes there is a need for the discovery of new proteases that can tolerate different conditions such as temperature and pH. So far protein engineering, mainly by random mutagenesis and directed evolution, have been used to improve the stability, activity or specificity of enzymes and today most proteases are genetically modified and/or produced by modified strains [67]. However, protein engineering does not cover the growing need for proteases with new characteristics, which has led to a shift in how enzyme discovery is performed and now a lot of focus lies in the development of new methods and in searching for novel enzymes in new habitats.

Enzyme discovery

The need for novel enzymes is large and as the applications of biocatalysts increase, the demand will increase with it. Historically, biocatalysts could only be obtained by isolating a microbe and growing it into pure-culture to isolate the enzyme [68], this is in part still the case today and the problem still prevails as approximately 99% of all microorganisms are still unculturable [69-72]. However, the term “unculturable” is somewhat misleading since the term simply mean that we are not yet able to obtain the microorganisms in pure cultures, not that they are actually unculturable [73]. Using
biocatalysts has many advantages, but there are limitations to the range of chemical reactions they can perform. As conditions needed for industrial processes can differ significantly from the physiological conditions they were evolved to perform in this can impair their optimal performance [74, 75]. This has, and still is, largely solved by using protein engineering to enhance stability and activity of the enzyme, a task that can be laborious and lacks rational design [24, 76]. With limited options of biocatalysts available, the industrial processes often have to be adapted to fit the enzyme instead, often leading to suboptimal processes and reaction conditions. Thus, to fit different process conditions, protein variants with different properties but of the same catalytic function is needed to cover a range of different process conditions. Given the large majority of uncultivated microbial diversity it is highly possible that the ideal catalyst already exists in nature [77, 78]. It is, thus, likely that there is a more efficient natural counterpart of all industrial enzymes currently in use [23].

The ideal biocatalyst should have high turnover rates i.e. be highly active and the selectivity of the enzyme has to be correct for the process. Some cases call for a promiscuous enzyme where specificity is low, as in the detergent industry and in some cases the specificity has to be high as in the production of active pharmaceutical ingredients (API) or fine chemicals. Stability under the specific conditions needed for the particular process is crucial, it effects the economics of the process in several ways. The lifetime of the enzyme determines how often it needs to be replenished, and decreased catalytic activity due to poor stability prolongs the process time [77, 79]. To find these possible ideal enzymes in the uncultivated microbial biomass there has been increased interest for more pure-culture independent methods for enzyme discovery [80]. There have been massive advancements in technology in the last decades providing tools such as, mass spectrometry for protein identification and next generation sequencing for genome sequencing that have evolved into new scientific areas. These tools have in turn advanced the field of enzyme discovery and made it possible to find new enzymes without having to rely on pure-culturing.

Major advances in DNA sequencing and bioinformatics, specifically next generation sequencing (NGS), allows for parallel sequencing on a massive scale. Sequencing genomes of whole microbial communities (metagenomes) has become available to a large part of the scientific community, largely due to the very reduced costs for sequencing [81]. This has evolved into the field of sequence-based metagenomics which is the random sequencing of the DNA extracted from a microbial community inhabiting a natural or engineered environment [82] and this is a well-established method for
enzyme discovery [83]. These major advancements in annotation of genomes of organisms and metagenomes along with better techniques for mass spectrometry for protein identification have also led to analysis of the protein complement of organisms (proteome) and microbial communities, evolving into the field of metaproteomics [84, 85]. These -omics fields have both been utilized and combined in the work of this thesis and their importance to enzyme discovery will be discussed in further detail below.

Metagenomics

The term metagenomics was first coined by Handelsman and co-workers in 1998 [86] when cloning the metagenome of an environmental soil sample to access the collective genomes and biosynthetic machinery of the microbial microflora. Microbes in nature exist in communities where the composition and dynamics are dependent on external factors such as pH, temperature and salinity. To study the entire collective genome of a microbial community is, however, not possible using conditions used when pure-culturing since this will only benefit microbes that are able to adapt to laboratory conditions [87]. Metagenomics is a field of pure-culture independent genomic analysis of microbial communities. It is largely used to address the pure-culturing problem and being able to access and study the genomes of microbes which would otherwise be unavailable [88]. In 1985 Pace and colleagues created a new branch of microbial ecology by analyzing 5S and 16S rRNA in environmental samples to describe the microbial diversity without pure-culturing. Early experiments where tedious since RNA had to be directly sequenced, but after PCR technology was introduced for amplification of almost entire genes the discovery of diverse taxa for habitats from all over the world was accelerated [89]. These earlier techniques based on PCR have been used for decades but have limitations regarding the resolution they can provide since the information detail needed to make deep description of whole complex microbial communities is challenging. By major advancements in the sequencing of whole genomes, especially by the development of next generation sequencing (NGS), large scale sequencing has revolutionized and simplified sequence library preparation by omitting the cloning step, also leading to a major drop in costs [90]. After sequencing, the reads can be assembled into progressively longer contiguous sequences, or contigs, and then finally a whole genome can be assembled [91]. Even though we have come very far since the first complete sequencing of a genome, this still poses some problems. When sequencing a single organism, assembly is easier and computational gene localization is possible since there is only one organism to consider. When an entire microbial community is studied
the assembly becomes more difficult, and it is more likely than not, that not every stretch
of the entire genome of every single species of microorganism present becomes
sequenced [91, 92]. Nevertheless, metagenomics have come a long way and the potential
for using the genomic information gained for searching the genomes for novel
biocatalysts is of great interest and potential [93]. Metagenomics have two different
approaches for screening the genomic data collected, sequence-driven (homology-based)
and function-driven (activity-based) analysis. The use of these approaches for the
discovery of novel enzymes will be further described below.

The sequence driven approach identifies genes based on homology in databases.
Classically it is performed using PCR-primers constructed from known conserved
regions, usually a catalytic site in enzymes. With the rapid technology of NGS it is now
possible to screen whole metagenomes using bioinformatic approaches, called “in silico
screening”, thus bypassing the need for laborious library construction [94]. A major
limitation using this analysis is that it is impossible to correctly annotate a novel gene or
function, since if it is truly novel it will not be present in a database [95, 96]. However,
by using this approach the dependency of gene expression in a host is alleviated [97].
Although several novel enzymes of industrial importance have been discovered using this method it is still primarily used for mapping functionality and metabolic pathways.

Function-based screening is the only approach which is useful for finding truly novel gene classes or functions since sequence information is not needed for comparative reasons, (figure 3) [98]. Libraries are constructed from the extracted genomes and cloned into a host. The size of the library created depend on the screening method but for function-driven analysis, where there is no need to characterize complex pathways involving many genes, a smaller library is sufficient for discovery of new metabolic functions or activities [99]. The function driven approach is mostly used in biotechnological studies but is limited by the fact that the genome or genetic library created has to be expressed in a host system. Several systems are being used in metagenomics but *Escherichia coli* is the preferred host [100]. The low “hit rate” is a problem in this type of screening, this is due to multiple factors such as the host-vector system, size of the target gene, it’s abundance in the metagenome, the assay method and the efficiency of gene expression in the host [101, 102]. The efficiency of the gene expression in the host is a big issue since it has been shown that e.g. *E. coli* is only able to express 40% of inserted foreign genes [98, 100]. To alleviate this problem several different host strains and vectors have been tested [94].

The use of metagenomics for discovery of novel enzymes is of great interest and is growing. There are several approaches for using metagenomics for enzyme discovery, as described above. However, there are, also still a lot of limitations, such as the very low “hit rates” making the work tedious, time consuming and costly and the gain often very low. Metagenomics also has the disadvantage of looking into the genome, thus, only studying the genetic potential, not which proteins the members of a microbial population actually express. That is, metagenomics does not provide any information about what genes are actually active under a given condition or metabolic need. Thus, by only extracting and analyzing the genome from microorganisms thriving in a habitat which have similarities to the conditions needed in industrial processes, some crucial information is not obtained. That is, simply because a gene coding for an enzyme of a certain activity is present and identified in a metagenome, does not mean that the gene is actually expressed and used under the conditions under which the metagenome was collected. To find the proteins that are expressed under certain conditions and needs we have to look at the proteome or more specifically the metaproteome. Metaproteomics is not yet used to the same extent as metagenomics when it comes to enzyme discovery but
has been partly used in the work presented in this thesis and is gaining interest and will be further described below.

Metaproteomics

The term “metaproteomics” was first used by Wilmes and Bond in 2004 [103] as “…the large scale characterization of the entire protein complement of environmental microbiota at a given point in time…”, when they used a proteomic approach for analyzing the proteins from a microbial community derived from activated sludge. Metaproteomics has in the same way as metagenomics evolved as an answer to the pure-culturing problem and to answer the questions where metagenomics falls short, of what the proteins the metagenome is expressing, and thus the actual functionality in relation to the metabolic pathways [104, 105]. The initial focus of metaproteomics was with the studies of the functionality of microbial communities [106, 107], this is still a main focus of this field today but it is also beginning to being used for other purposes such as finding novel enzymes for industrial purposes [108] which is the scope of this thesis.

A limitation of metaproteomics is the large quantity of non-proteinaceous substances that follow with complete environmental samples such as humic acids and other residues depending on the environment [109]. This affects the protein extraction [110] and it is not possible to use a standard extraction method for all samples. The protein extraction method has a large impact on the number of proteins identified downstream and is a bottleneck in the proteomics field. Thus, using several different extraction methods is something to be considered. Since proteins cannot be amplified like DNA using PCR, the extraction method is of major importance for the amount of protein recovered for downstream analysis. There is also a lack of genomic data, which is needed for identification and annotation of the proteins. Since the publishing of metagenomic data from sampling in the Saragasso sea by Venter et.al 2004 [111] and the rapid development in sequencing technology, the number of sequenced genomes has increased dramatically [84]. However, lack of genomic data can still be a problem when studying a microbial community from which this data is not available for comparison. The availability of high quality genomic sequence data of the studied genome allows for tailored databases to be made, improving the number of proteins possible to identify [112].

The metaproteomic workflow starts with protein extraction and purification often followed by a separation of the proteins. This can be done using gel-based separation, such as two-dimensional gel electrophoresis, where there is separation of full-length
proteins based on net charge followed by a separation by size, or separation can be done using two-dimensional liquid chromatography (2-D LC) separating the proteins in a liquid phase by using two different columns with different separation conditions in succession. Proteins can then be directly analyzed by mass spectrometry giving information on isoforms and posttranslational modifications and with the mass and fragmentation an identity can be obtained. This approach is not useful for complex samples such as in metaproteomics since the mass prediction is suggestive, coupled with the many modifications possible that can alter the mass of a protein and trouble detecting larger proteins (>50 kDa) it is not a reliable for identification. This is called a top-down approach. The most common approach for metaproteomics is the bottoms-up approach and has come to be called shotgun-proteomics. All the extracted proteins are digested into peptides most commonly using trypsin, and are separated using techniques such as, online-LC coupled with MS/MS or multi-dimensional LC followed by MS/MS, and identification of the proteins using the peptide fragments or de novo sequencing of the peptides to search against protein databases. When using gel-based separation, by 2-D gel electrophoresis, as has been used in the experiments in this thesis, it is not a straightforward shotgun approach since single proteins are first selected from the gel and then cleaved by trypsin, unlike digesting all the proteins recovered from extraction. 2-D (two-dimensional) gels can be used when there are several different conditions affecting the microbial community to investigate differences in expression, as seen in [113]. Identifying a single peptide confidently back to a single protein in a whole metaproteome or a single protein in a huge protein database such as NCBI isr has some problems, many peptides are homologs or redundant and can be assigned to many proteins thus lowering the fidelity of identification. By first using 2-D gel electrophoresis for separation of full length proteins there should theoretically only be one protein per sample analyzed on MS/MS, that along with knowing the mass and pI of the protein identification should be somewhat facilitated [114].

Metaproteomics was initially largely gel-based using one dimensional gel electrophoresis (1-DE) and two-dimensional gel electrophoresis (2-DE) to separate and visualize the proteome. This required one gel per sample and comparing the gels against one another which was difficult with large gel-to-gel differences and has now largely moved onto using 2-D difference gel electrophoresis (DIGE) where samples can be run together and only one gel is needed. There are several limitations using gels for separation, the resolution is not the best, it can be hard to differentiate between proteins on the gel, there is a risk of contamination of the gel by introducing keratins from the environment in laboratory and it is hard to pick protein-spots if this is done manually.
since they can be very small. To circumvent these problems, it is recommended to work in a clean lab, sterile if possible, to avoid contaminations, and to pick spots more accurately a picking-robot can be used. However, these are very expensive and not affordable in comparison to e.g. using upstream LC. Nevertheless, 2-D DIGE allow for separation not only based on pI and mass, but it is also possible to quantitate the relative amount of full-length protein in each protein spot on the gel which is not possible using merely LC for separation. It has been a shift where gels are being used less and separation is done directly coupled to mass spectrometry (MS) as a result of the evolvement of LC-MS/MS. Peptides from MS are analyzed for homology against a protein database and possible annotation is made [115]. To be able to correctly identify and annotate the proteins, the genomic data of the population investigated, or similar populations is needed, as advances in metagenomics enables more microbial genomic data to be collected it also help bring the field metaproteomics forward.

Metaproteomics has the potential to be used for enzyme discovery. However, this is currently not being applied to the same extent as the field of metagenomics but is gaining interest in the search for novel enzymes [108]. Metaproteomics alongside metagenomics has been used for novel protein identification in the work covered in this thesis. The combination of the two fields is gaining interest and has been named “metaproteogenomics”. This field and possible applications in enzyme discovery will be further described below.

Combining meta-omics techniques into metaproteogenomics

The concept of “proteogenomics” was introduced by Jaffe et. al in 2004 [116] to unite the potential of proteomics with global genome annotation. The characterization of a complete proteome is done by comparing protein or peptide data to a reference protein sequence derived from a genome database of the same sample. This need for both genomics and proteomics in combination for large-scale characterization is, thus, the origin to the expression “proteogenomics” [114]. When examining data from a microbial community the corresponding term is, thus, metaproteogenomics. Notably though, when the term is searched for in Web of Science only 16 publications are retrieved, showing that the field is in its very infancy. Annotation/identification of peptides/proteins using data from tandem MS (MS/MS) proteomics experiments is normally done against existing protein databases such as Swiss-Prot which is non-redundant and where the proteins are experimentally verified. Alternatively, against NCBI non-redundant database containing all entries for protein sequences derived from many NCBI resources
such as NCBI RefSeq, GenBank, PDB and Swiss-Prot [117]. However, when analyzing microbial communities with a large fraction of earlier not sequenced or analyzed microorganisms by using these public databases it is likely that a comprehensive representation of a specific sample will not be found, thus, hindering genomic annotation [118]. Proteogenomics is used mainly for genome annotation which is structural annotation, as in mapping genes, promoter and regulatory elements, and functional annotation which is understanding of the functionality of the genes. Metaproteomic approaches provides MS/MS data of the collected expressed genes (proteins) verifying protein-coding genes in the genome, it can also be beneficial for identifying missed protein-coding genes, confirm splice variant in eukaryotic genomes and correct overestimated protein-coding potentials [119].

For enzyme discovery in microbial communities in which the microorganisms cannot be pure-cultured it is crucial to obtain the full sequence of the protein for cloning of the gene, since the original organism, of course, cannot be pure-cultured and used for large-scale production of the identified enzyme. Finding a homologue will only tell you if the protein is of interest based on its function against similar proteins and does not provide the necessary information for cloning of the actual sequence coding for the identified protein. By using NGS it is possible to sequence whole genomes of microbial communities and by using bioinformatics it is possible to construct a database of hypothetical genes/proteins of the community investigated. Having a database containing the complete genome of the microbial community that is actually investigated to search against greatly improves the odds of finding the sequence of interest. The genomic database can be translated on all six frames and can be used for identification of novel enzymes using the peptide sequence data from the MS/MS analysis and thus providing the complete and correct protein sequence.

Biogas producing microbial community

Biogas consist of methane and carbon dioxide and is the end product in anaerobic digestion. It is produced in a multi-step process performed by a concert of microbes acting together under anoxic conditions. Biogas producing microbial communities can be found all over the world in a range of different habitats, such as the termite gut, the cow rumen, underwater thermal vents, landfills for waste, and anaerobic digestors. With the climate goals and fossil fuels being a finite energy source and contributor to climate change there is a call for renewable energy sources. Biomass is considered one of the most important renewable energy sources and biogas production is considered a
key technology in utilizing agricultural biomass for production of energy and fuel. Biogas is in many ways a better alternative for biofuels production than many other biofuels since it can be made from a variety of waste resources such as sewage sludge, manure, crop waste, food waste etc. In addition, the residues after anaerobic digestion can be used as valuable fertilizer for agricultural crops [120, 121].

Anaerobic digestion is a multistep process and is considered as four steps starting with hydrolysis, followed by acidogenesis (primary fermentation), acetogenesis (secondary fermentation) and methanogenesis, (figure 4). Hydrolysis is the first step where the microbes secrete extracellular enzymes to hydrolyze polymeric substances such as polysaccharides, proteins, nucleic acids and fats/oils to soluble monomeric organic materials which can be transported across the cell wall. These molecules are broken down in acidogenesis to volatile fatty acids (VFA) and long chain fatty acids (LCFA), alcohols, \( \text{CO}_2 \), \( \text{NH}_4^+ \) and \( \text{H}_2 \). In the second fermentation step, acetogenesis, the intermediate fatty acids and alcohols are turned into acetate, \( \text{CO}_2 \) and \( \text{H}_2 \) which are the precursors for the methanogenesis. In the final step two groups of methanogens act, group one (acetoclastic methanogens) split acetate into carbon dioxide and methane and the other group (hydrogenotrophic methanogens) uses hydrogen as an electron donor and carbon dioxide as an electron acceptor to produce methane [122-124]. With hydrolysis being rate limiting for many substrates there is a need for novel efficient hydrolytic enzymes that can be used to enhance the production in industrial biogas plants.

The biogas producing microbial community naturally secretes extracellular enzymes into its environment, making it an excellent candidate for enzyme discovery since extracellular enzymes are naturally adapted to withstand the conditions of the extracellular environment independently of the cell. By identifying hydrolytic enzymes secreted by the microbial anaerobic community it could be possible to improve the biogas production by adding these enzymes which are already evolutionarily adapted to withstand the conditions residing in an anaerobic digestor. In addition, the microbial community responsible for anaerobic digestion in biogas reactors serves as a very good model for anaerobic environments which is important because most extreme environments harboring extremophilic microorganisms are anoxic. Identifying enzymes from a biogas producing microbial community has been part of the work which will be presented in this thesis.
Metaproteogenomics for discovery of novel enzymes

By using metaproteogenomics we have developed a process for enzyme discovery which will be presented in this thesis. We have been able to utilize the advantages of metaproteomics for extracting proteins and retrieving protein data. Metagenomics have been used to create a translated metagenomic database of the whole collected metagenome of our model populations consisting of biogas producing microbial communities existing under laboratory conditions and kept in a metabolic steady state. We have with the metaproteogenomic method developed during the work of this thesis shown that it is possible to combine techniques from metagenomics and metaproteomics for the purpose of targeted enzyme discovery.

Figure 4. Biogas production can be divided into four steps, starting with hydrolysis of complex substrates followed by fermentation (acidogenesis) forming intermediary products such as alcohols, organic acids, ammonium and carbon dioxide, a second fermentation follows turning intermediaries into acetic acid and hydrogen which are utilized in the final step, methanogenesis, producing methane and carbon dioxide i.e. biogas.
Biogas producing microbial populations is of interest because of its natural secretion of hydrolytic enzymes which are of great industrial interest and there are many of these populations living under a variety of conditions with a very different microbial composition. This makes them highly interesting as targets for enzyme discovery. In paper I [125], it was shown that the conditions for an efficient and targeted enzyme discovery by metaproteomics are conceivable. In particular it was shown that it is possible to establish a viable microbial community using a chemically defined medium, which is necessary for the downstream analysis. By maintaining the microbial community on a chemically defined medium over time in a bioreactor, a physical and metabolic steady state will be reached. Thus, providing a stable condition for a base-line expression of proteins. It will further provide conditions for sampling of the extracellular metaproteome since the extracellular environment will be clean and only contain microbial metabolites and proteins produced by the actual community. In addition, since the nutrients (amino acids, sugars, fatty acids etc.) were provided as the monomeric components of the corresponding macronutrients (proteins, polysaccharides, triglycerides) the microbial incentive and need for producing hydrolytic enzymes was eliminated, leading to a low base-line expression of the analyzed enzyme activities. Most importantly, the expression of a specified and targeted enzyme activity could then be distinctively and strongly increased by inducing the expression and activity by exchanging the monomeric nutrients for the complex biomacromolecule. Thus, to provide conditions where a certain enzyme activity is turned “off”, and another condition where the same enzyme activity is turned “on”, while other activities remain at base-line level. Thus, opening up for the metaproteomic sampling of two distinct conditions in which it is mostly the expression of the targeted enzyme activity that differs between the two samples.

In metaproteomic analysis protein sampling and extraction methods are a major limitation since it is only the proteins that are actually picked up in protein sampling that have a chance to become analyzed and identified. That is, simply because a protein is not identified, that does not necessarily mean that the protein is not present in the environment studied, it could simply be that it is not picked up by the protein sampling/extraction method employed. Another problem is that protein sampling and extraction of samples collected at natural environmental sites have a low reproducibility and, hence, large variability between samples, which can lead to the misinterpretation of metaproteomic data. That is, the differences originating from sampling/extraction could be interpreted as false differential expression between two sampling time points.
Figure 5. Schematics showing the metaproteogenomic approach for discovery of novel enzymes used during the work presented in this thesis. Sample is taken from a microbial community kept at enzyme suppressed metabolic "steady state". The sample is divided into a control, fed with the same readily available nutrients as in the biogas reactor, and an induced sample, which has the readily available nitrogen source replaced by a complex source (BSA). Enzyme activity and biogas production is monitored, and samples are collected when enzyme activity is high in the induced sample. Extracellular proteins are extracted from the extracellular fluid of both samples and compared by 2-D DIGE. Protein spots identified as upregulated are picked and in-gel digested with trypsin and further analyzed by MS/MS. Generated de novo peptide sequence tags are searched against the sequenced metagenome, translated into a hypothetical protein database, from the original sample. Identified complete open reading frames are annotated using BLASTp to search against the NCBInr protein database.
This is probably the reason that one of the main benefits of metaproteomics, i.e. to analyze the difference in protein expression at different time points, is often not exploited. In addition, because of the presence of many contaminating substances in the extracellular environment that can negatively influence the proteins separation in 2-D gel electrophoresis or 2-D Nano-LC/MS the extracellular metaproteome is seldom analyzed. Rather, cells are collected, washed and lysed before the intracellular metaproteome is extracted and analyzed. Thereby overlooking the actively secreted extracellular hydrolytic enzymes that are often the target of enzyme discovery for biotechnological applications. Furthermore, since conditions between environments can to a large extent differ, there is no single standard method that can be used for all samples, although certain protein extraction methods, such as TCA precipitation, is used more often than others. Ultimately, for each specific sample many extraction methods need to be tried to find the best method that provides samples that are clean, picks up the largest amount of proteins, and a representative collection of proteins of the sample. In paper II [126], several extraction and precipitation methods were examined to determine the best method for the sampling of the extracellular fraction from the anaerobic biogas producing microbial population. There is no optimal sample preparation model and ultimately there will always be variations in the amount of protein extracted and which proteins are able to be extracted using a certain method.

The objective of running a bioreactor of a full microbial community was to expand the microbial source for targeted enzyme discovery, from pure-cultured microorganisms to a full microbial community. Furthermore, since this was accomplished in the unconventional way of using a chemically defined medium with monomeric nutrients, the community structure needed to be analyzed to verify that microbial communities maintained under such conditions still has a high species richness and diversity and was composed of a representative selection of microorganisms at the conditions created.

For this purpose, in paper III, the methanogenic communities of a mesophilic and a thermophilic bioreactor maintained on the same chemically defined medium were analyzed by shotgun next generation sequencing metagenomics. In addition, the sequencing data was then assembled and translated on all six frames and to construct a database of hypothetical proteins, derived from the very same communities. By applying all of the above, targeted enzyme discovery of extracellular proteins in full microbial communities is made possible, (figure 5). That is, as described in paper IV, by controlling the gene expression of a targeted enzyme activity in a full microbial community, extracting and comparing the gene expression of extracellular proteins
between an induced sample and a non-induced reference by 2-D gel electrophoresis, and then analyze the *de novo* peptide sequence tags of proteins identified as up-regulated against a database of hypothetical proteins, derived from the actual microbial community. By this the full and correct sequence for a protein identified as upregulated will be obtained and the function of the protein can be retrieved from analyzing it against public databases such as NCBI nr. Thus, giving information on both the function and the full-length protein sequence, which is crucial for cloning and production of any targeted and novel enzyme of interest.
Aims of the study

The aim of this thesis has been to use metaproteomic, metagenomic and bioinformatic tools and techniques to develop a method for targeted discovery of novel enzymes that can be of industrial interest.

Aims of the papers:

Paper I  A biogas producing microbial community was adapted to and kept at metabolic steady state by using a chemically defined medium with monomeric nutrients. This ensures that enzyme production for hydrolysis of nutrients was not necessary and the production of extracellular hydrolytic enzymes was suppressed. The aim was to be able to induce specific enzyme production, at will, in the microbial community in response to the addition of complex carbon or nitrogen nutrients.

Paper II  Sample preparation is the most crucial stage in the proteomics workflow and it determines how much of the complete proteome is available for downstream analysis. The aim was to assess different protein extraction and precipitation methods and to determine which would be most suitable for extracellular proteins in general, and from the anaerobic microbial community used as a model system in particular.
Paper III  Shotgun metagenomic analysis was performed on biogas producing microbial communities maintained at metabolic steady state at mesophilic and thermophilic conditions using next generation sequencing. The aim was to sequence the metagenome as completely as possible and to investigate how well the communities represented a natural microbial community of anaerobic digestion and further to investigate if the communities were representative of the two temperatures. Thus, to validate that the communities were complete and representative communities for enzyme discovery despite of the unconventional condition of feeding with a chemically defined medium of monomeric substrates.

Paper IV  Theories of targeted induction of protein expression from paper I was exploited together with an extraction method from paper II. Furthermore, shotgun sequence data from paper III was used to construct a database of hypothetical proteins. The aim was to induce a targeted protease response in two thermally different biogas producing microbial communities to be able to discriminate the up-regulated proteins from the large background of constitutively expressed proteins and identify not only the function of extracellular up-regulated proteins, but also the complete full-length sequence for these proteins.
Biogas producing microbial communities at metabolic steady-state

Two anaerobic biogas producing microbial communities were used in this study. One mesophilic maintained at 38 °C, which was started from a full-scale biogas reactor run on mainly slaughter house waste, and one thermophilic community maintained at 52 °C and started from thermophilic a co-digestion plant. Digestates were collected from both plants and transferred to two separate experimental bioreactors, (figure 6) [127]. The temperature was automatically regulated within ±1 °C via a temperature probe inside the bioreactor, a thermostat and a heating blanket. The bioreactors were continuously mixed with a stirrer and the biogas production and methane concentration were monitored continuously. The microbial communities were fed a chemically defined medium which contained more than 40 different substances and were based on several other recipes [128-130]. The recipe was modified a number of times until a final composition was found that enabled the microbial communities to reach metabolic steady state. The reactors and the defined medium is describe in detail in paper I [125]. By changing the chemical composition of the medium over a time period of about five months it allowed the microbial communities to adapt to the changes and maintain a complete and viable biogas producing community. It also allowed for any contaminants and substances from the original digestate sample to be either completely consumed or
washed out. All nutrients fed to the biogas reactors were in monomeric form, meaning that the microbial communities had nutrients readily available and thus, the microorganisms had no need to produce hydrolytic enzymes in order to make nutrients available.

**Figure 6.** Experimental biogas reactor for biogas producing microbial community existing under metabolic steady state. Image above depict the mesophilic population being kept at 38 °C and anaerobic conditions with constant stirring.

**Induction of a targeted enzyme expression in microbial communities at metabolic steady state**

By maintaining the microbial communities under artificial conditions, at metabolic steady state with no or low hydrolytic activity, a unique starting point was established when it comes to enzyme discovery. By altering the composition of the defined medium or by adding or removing nutrients the metabolic equilibrium maintained by the microbial population could be altered. It has been shown that microbial enzyme production is dependent on the nutrients available to the microbes [131, 132]. If there is a deficit in readily available nutrients, and more complex substrates are present, the microbes will secrete enzymes in its surroundings that will hydrolyze proteinaceous, fatty or carbohydrate containing substances into suitable monomeric nutrients that can be used by the microbes until the demand is met [133]. Thus, deficiency in readily available nutrients in combination with the availability of nutrients in complex form induces the response of enzyme production [134]. These principles have been used previously by our
group to induce cellulases and [108, 135] and have been utilized in paper I [125] and for the induction of proteases in paper IV with the purpose of discovering new proteases which is the main scope of the work presented in this thesis. For protease induction in the biogas producing microbial communities, sampling was done just before daily feeding so that the majority of all readily available nutrients would have been consumed. The samples where then divided into a reference that was fed with exactly the same medium as in the bioreactor and at the same organic loading rate (OLR), and an induced sample which received the same nutrients as in the reactor apart from any nitrogen containing substances (i.e. no amino acids or NH₄Cl) creating a demand for nitrogen in the microbial community. The nitrogen source was instead supplied as protease free bovine serum albumin (BSA) which is a complex nutrient source. By removing readily available nutrients and replacing it with a complex nutrient a targeted response from the microbial population was induced. This response in secreted extracellular protease activity can then be measured against the reference which have all the activities of the induced sample in the extracellular liquid, except for the hydrolytic activity. Thus, the proteins which are upregulated in the induced sample can be differentiated from the large background of proteins that are constitutively expressed in both samples, by comparing the induced sample with the reference sample using gel-based proteomic techniques which will be described further on.

**Enzyme assay**

When enzymes are present in a sample it is possible to measure the activity by introducing a substrate that the enzyme will recognize and process, the substrate can sometimes be synthetic and is often coupled to a molecule which can be measured fluorometrically, colorimetrically or by luminescence. In fluorometric substrates the close vicinity of many fluorescent molecules in a substrate renders the fluorescence quenched, so if there is no enzyme present to digest the substrate there is no measured activity. When a hydrolytic enzyme binds to and hydrolyses the substrate the molecule is digested into smaller pieces causing the quenching to disappear and it will fluoresce or induce a visible color change that can be measured spectrophotometrically. This gives a relative quantifiable estimate of the amount of enzyme present in the sample.

Enzyme activity is since 1978 defined as the conversion of one mole substrate to product per second in a specified assay system. The unit is *katal* (kat). It is still very common to use the old enzyme unit (U) which relates to *katal* as follows: 1 U catalyzes a rate of 1 µmol/min = 1/60 µmol/s ≈ 16.67 nmol/s; 16.67 nkat catalyze at a rate of 16.67 nmol/s.
and therefore 1 U corresponds to 16.67 nkat [136] according to the Nomenclature Committee of the International Union of Biochemistry (NC-IUB).

During the work described in this thesis enzyme assays measuring the release of the dye resorufin by hydrolysis was used. Resorufin is a fluorescent dye but can also be measured colorimetrically. Resorufin is attached to a specific substrate depending on the target enzyme which during the work in this thesis was either cellulases or proteases.

Protease assay

Proteases are enzymes which hydrolyze peptide bonds. The “Universal protease substrate” from Roche, Germany, was used. This is casein proteins labeled with resorufin, (figure 7). Upon protease activity the substrate is cleaved and peptides with resorufin becomes solubilized. After digestion, undigested resorufin-labeled casein substrate is precipitated with TCA and the remaining non-precipitated solubilized resorufin-labeled peptides can be detected by absorbance measurement at 574 nm. Exact detail on the assay can be found in paper I [125] and paper IV. A universal protease substrate was chosen since the proteases in the sample was unknown and most likely have many different specific activities. It is important to be aware that although many proteases cleave casein there are some activities that cannot be detected using casein. However, this would be the case with any assay chosen.

![Figure 7. Universal protease substrate. Casein labeled with resorufin, used as substrate in protease assay. Resorufin-labeled peptides are released upon proteolytic activity and the absorbance or fluorescence can be measured.](image-url)
Cellulase assay

Cellulases cleave the polysaccharide cellulose by hydrolysis. The assay used is the Fluorescent Cellulase Assay Kit from Marker Gene Technologies, Inc. USA. It uses a resorufin-labeled disaccharide cellobiose as substrate, (figure 8), upon hydrolytic cleavage of the $\beta$-1,4 glycosidic bond resorufin is released into the solution and can be detected by measuring the fluorescence emission at 595 nm. The fluorescence is followed over time and is compared to a standard curve of known dye concentration. The release of resorufin is measured in $\mu$mol min$^{-1}$ and the relative cellulase activity can thus be calculated.

![Resorufin cellobioside](image)

**Figure 8.** Resorufin cellobioside. Used as substrate in cellulase assay. Hydrolytic cellulase activity releases resorufin and the increase in fluorescence can be followed over time.

Protein extraction and precipitation

Protein extraction and precipitation is the crucial first step, after sampling, in the metaproteomic workflow and determines what part of the metaproteome will be accessible for analysis irrespective of downstream methods. In part, the importance of protein extraction comes from the fact that there is no amplification process available for proteins like there is for DNA by PCR. There is further no single protein preparation method that can be expected to work for all types of samples. Thus, it is more likely that a method has to be empirically found or developed for each new environment investigated [137].

Enzyme concentration is often low in environmental samples and is why enzyme extraction and precipitation is important not only as a way to remove contaminants, but also for concentrating the proteins. This is something that was of even larger importance during the work of this thesis, where the focus was on extracellular hydrolytic enzymes found in the medium. That is, the proteins targeted in this work were much more diluted
in comparison to when studying intracellular samples, which can first be concentrated by pelleting and washing the cells before lysis. There are a number of different ways to prepare proteins for metaproteomics. The most common way is to precipitate the proteins from the sample, either with or without a preceding extraction step. Proteins possess both ionizable and hydrophobic groups which interact with the solvent. This can be used to make proteins agglomerate and finally precipitate [138]. Some precipitation methods will be briefly described below.

Precipitation methods

Precipitation with salts is most commonly done with ammonium sulfate, the salt changes the electrostatic forces responsible for solubility by acting on the water molecules surrounding the protein. Precipitation by organic solvents lowers the dielectric constant of the solution displacing the water surrounding the protein leading to a decrease in solubility and precipitation. Commonly two-carbon solvents such as ethanol and acetone are used [139]. Precipitation by polymers are usually done using polyethyleneimines and polyethylene glycols (PEG) and acts by affecting the solubility in a similar fashion as organic solvent by mechanically/physically and in a thermodynamic sense push the proteins out of solution. Isoionic precipitation is precipitation at the proteins isoelectric point (pI) where the net charge is zero and the proteins solubility is markedly minimal [138, 140]. This method cannot be used on metaproteomic samples since there are a variety of proteins with different pI. In precipitation by pH-denaturation the extreme pH causes denaturation due to sensitive areas in the protein structure obtaining similar charges which cause internal repulsion and possibly also affect weak charges which was previously important for holding the protein structure together [141]. TCA is the most commonly used in precipitation by lowering the pH but its efficiency depends on the physiochemical properties of the proteins in the sample [142].

Sometimes an initial extraction step, e.g. phenol extraction, is efficient for removing contaminants before precipitation from samples with low concentration. Phenol extraction was first used to remove proteins from DNA samples together with chloroform [143, 144] but has evolved to also being used for protein extraction [145, 146]. Phenol dissolves proteins and lipids leaving water soluble substances such as nucleic acids in the aqueous phase [147]. When studying different sample preparation methods in the work in this thesis, extraction methods using phenol, followed by precipitation in methanol were found to be the most beneficial for our samples and was later used. However, also other sample preparation methods showed promising results. [126]
2-D gel electrophoresis

2-D gel electrophoresis (2-DE) is a separation method for protein samples and was developed by O’Farrel in 1975 [148]. The separation is performed in two stages, the first is isoelectric focusing (IEF) and separates the protein according to their isoelectric point (pI) and the second dimension is sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) separating the proteins according to size. When 2-DE was first introduced there was no -omics fields, not even genomics. Protein identification was quite problematic since there was no mass spectrometry and protein identification was performed by co-migration with purified known proteins or blotting with antibodies. IEF was done with unsupported tube gels with 4% acrylamide which caused a lot of issues regarding reproducibility. Another problem was that the pH gradient was established using carrier ampholytes which was unpredictable and affected the reproducibility. There was a major shift in the late 80s when immobilized pH-gradients were introduced eliminating the use of carrier ampholytes. The integration of the pH gradient with the SDS-gel in the form of tube gels or gel slabs were not successful and finally the solution that is still used to day was introduced, immobilized pH gradients (IPGs) fixed in a gel strip on a plastic base. This made the technique much easier to use and reproducibility was enormously improved. 2-DE and the possibility to be able to analyze single spots of interest from the gel was one of the starting point of proteomics [149]. However, the first and a key step for successful 2-DE is still sample preparation.

2-DE make it possible to compare samples from different conditions, e.g. a control sample and a treated sample [150]. By comparing the gels using image analysis it is possible to identify proteins of interest that are either up- or downregulated. The proteins of interest can be picked from the gel and further analyzed using MS. This approach was used in this study although we used a technique called difference gel electrophoresis (DIGE) which enabled us to have the reference and induced sample on the same gel, thereby avoiding gel-to-gel differences. DIGE with be described in detail further on.

Sample preparation for two-dimensional gel electrophoresis

Sample preparation for 2-DE is done with the aim of converting the protein sample into a physiochemical state suitable for IEF, while still preserving the native charge and molecular weight (MW) of the proteins. In most cases this means a need for solubilization, disaggregation, denaturation and reduction of the precipitated proteins. This is only the case if the IEF does not need to be run under native conditions. However, most 2-DE strategies is to resolve as many proteins as possible within a certain pI/MW-range to facilitate the comparison of two set of data samples and it is preferable to do this
under denaturing and reducing conditions [151]. Solubilization buffers usually contain chaotropes (urea and thiourea), nonionic and zwitterionic detergents (Triton X-100 or CHAPS), reducing agent (DTT or DTE) and carrier ampholytes. Chaotropes, urea or a combination of urea and thiourea are usually used in high concentrations as they are good at disrupting hydrogen bonds causing denaturation. Their uncharged state ensures they do not affect the charges of the proteins and make them perfect for IEF. Ionic and/or zwitterionic detergents is included in the buffer to prevent hydrophobic interactions between hydrophobic domains and thereby avoiding loss of protein due to aggregation and precipitation. Reducing agents prevent reformation of disulfide bonds and ensure complete unfolding of the proteins and maintain the proteins in their reduced states. Carrier ampholytes aid in the solubility of the proteins as well as the prevention of carbamylation of the proteins amine groups which is a risk when using urea. Urea in solution exist in equilibrium with ammonium (iso)cyanate which can react with the α-amino groups at the N-terminus as well as the ε-amino groups of lysine residues which causes artifacts blocking the N-terminus and alterations of pI-values of the proteins. To avoid carbamylation temperatures over 37 °C should be avoided and carrier ampholytes are included in the solubilization buffer [152, 153].

Isoelectric focusing (IEF) – first-dimension separation

In the first-dimension separation, the proteins are separated based only on their isoelectric point (pI) which is the pH where the net charge of a protein is zero. The pI of a protein is determined by the number and type of acidic and basic amino acid residues they contain. IEF is carried out under electric current, i.e. electrophoresis, in a pH-gradient which allows each protein to migrate to its isoelectric point where the net-charge is zero, (figure 9). This is achieved since the charge density of the proteins decrease the closer they are to their isoelectric point and the protein stop migrating when their net charge density reaches zero. Diffusion is a factor but when proteins move away from their pI they become charged and migrate back to their pI. There may also be an initial sieving effect where the proteins migrate based on their size, but running the gel for a long time ensures size-independent separation [154].

The pH-gradient is chosen depending on the proteins of interest. The pH gradient is immobilized on an IPG-strip which is a gel on plastic back support and can be bought pre-made. The gel on the IPG-strip has to be rehydrated using the same conditions as the sample buffer i.e. pH and ionic strength. The proteins sample can be loaded onto the IPG-gel strip in a number of different ways during or after the rehydration of the IPG-strip. In the work presented in this thesis we have chosen anodic cup loading where a small cup
is placed on top of the IPG-strip at the anodic side and the sample is placed in the cup. This, however, puts restrictions on the sample volume [155] and the protein sample need to be of sufficient concentration. The IPG-strip is placed in an IEF cell and electrodes are placed on the gel connecting the low-pH side of the gel to the anode and the high-pH side to the cathode. The proteins enter the gel under a moderate field strength which also allows for salts remaining in the sample to migrate slowly to their final positions as to not disrupt the gel and the protein migration [155]. The voltage is increased in steps until the high field strength needed for isoelectric focusing is reached.

**Figure 9.** Isoelectric focusing (IEF). Protein are separated according to their pI in a fixed pH gradient. Proteins migrate under an electric current until they reach the pH where their net charge is zero. IEF is the first-dimension separation in 2-D gel electrophoresis.

**SDS-PAGE – second-dimension separation**

SDS-PAGE is the second-dimension separation in 2-D gel electrophoresis, (figure 10). It separates proteins based only on molecular weight (MW). Proteins are separated in a gel made of acrylamide monomers and the cross-linker N,N’-methylenebis-acrylamide, and by varying the concentrations of these gels with different pore sizes can be created. That is, acrylamide polymerizes into polyacrylamide chains forming a mesh cross-linked together with N,N’-methylenebis-acrylamide [154].

The basis of the technique is in exposing the proteins to the negatively charged molecules of sodium dodecylsulfate (SDS). Proteins are most commonly run in denatured mode and SDS binds to the polypeptide chain of the proteins stoichiometrically, the large charge of tens or hundreds of SDS-molecules thereby dwarfing the internal charge of the proteins. Larger proteins have a longer polypeptide chain and are thus able to bind more SDS-molecules and thus being more negatively charged than smaller proteins [154]. However, the charge to size ratio will be the same for all proteins and, thus, the proteins will migrate according to size and not charge. DTT and β-mercaptoethanol are usually added to break disulfide bonds to reduce the protein to its primary structure.
The sample added to 2-DE is in the IPG-strip after IEF and the proteins are first reduced by incubation in a sample buffer containing SDS with addition of DTT and in a second step incubated in sample buffer with the addition of iodine acetamide which alkylates the cysteines on the reduced disulfide bonds ensuring that re-oxidation and reformation of disulfide bonds does not occur. The IPG-strip is added close to the gel and by encapsulating it in molten agarose it ensures that the strip is sealed to the gel. The proteins enter the gel during a step with lower field strength and are then run in one continuous step with higher field strength until adequate separation is reached. The IPG-strip is separated from the gel and the proteins are fixed in the gel by being incubated with a mixture of ethanol and acetic acid. The proteins are then visualized using dyes. There are a number of different dyes such as Coomassie, silver stain and fluorescent dyes such as SYPRO which differs in their detection sensitivity.

2-D DIGE

2-D difference gel electrophoresis (DIGE) was introduced by Ünlü et. al. in 1997 [156] as a modification of 2-DE and a solution to gel to gel differences by excluding the need to use one gel per sample. By labeling the two samples with different fluorescent cyanine dyes called Cy3 and Cy5, which are N-hydroxyl-succinimidyl esters (NHS) and covalently bind to the ε-amine groups of lysine [157], they could be run together on the same gel and then images could be obtained using different wavelength filters to differentiate between the samples. The images could then be superimposed on each other.

Figure 10. 2-D gel electrophoresis. The second separation dimension in 2-D gel electrophoresis is SDS-PAGE which separates proteins according to size. Proteins migrate in the polyacrylamide gel under an electric current. The gel is a reverse sieve and larger proteins migrate slower.
and differences in expression are easier to detect. When using one gel per sample there is always gel-to-gel differences which cannot be avoided [158] and much focus was put into computationally solving this using image analysis. However, the computational solutions were limited in their success due to the many reasons of gel-to-gel differences and the number of variations possible [159]. When the third dye, Cy2, was introduced it opened up new possibilities for experiment setups. The third dye was used for a pooled sample containing an equal amount of all samples included in the experiment. This sample is run on the gel with the Cy3- and Cy5-labeled samples and acts as an internal standard which can be used to normalize against for all spots on all gels [160, 161]. When running DIGE experiments it is common to run two gels and do a dye-swap between the Cy3- and Cy5-labeled samples to reduce any possible dye affinity differences that might exist in one of the samples. The Cy2-labeled sample further reduce gel-to-gel differences by acting as an internal standard, (figure 11).

Figure 11. Scheme for running DIGE using minimal labeling.
When labeling the proteins, it is important that the dyes do not affect the charge of the proteins, also the dyes must have similar weight and react with the same amino acid as well as having their own distinct fluorescent characteristics. It is important that the labeling does not disrupt the proteins migration during 2-D gel electrophoresis. This is unavoidable with covalent modifications and there is a small change in the migration during the mass separation. Two different types of labeling were found to be acceptable to avoid this problem, minimal labeling and saturation labeling. In minimal labeling each protein is only labeled by one dye molecule. Minimal labeling, which was used during the work presented in this thesis, targets lysine residues and to ensure that, on average, only one lysine per protein is labeled, conditions are used to ensure that only about one in 20 proteins has one lysine labeled and the rest of the proteins remain unlabeled [162]. In saturation labeling all possible residues of one type is dye-coupled. Saturation labeling targets cysteines since there is generally fewer of them per protein and the dyes are soluble zwitterions aiding in complete labeling of all cysteines. Saturation labeling is highly sensitive and successful when working with low-concentration samples [162]. Since DIGE has a restriction on the amount of protein that can be run on the gel there can be a problem with the spot detection for picking the spots for MS/MS analysis, this also requires post-staining. To circumvent these problems a preparative gel can be prepared containing a higher amount of protein and no labeling affecting the spot identification.

The work presented in this thesis used DIGE to detect expression differences of proteases between an enzyme induced sample and a reference from biogas producing microbial communities (paper IV).

**Image analysis**

An image of the proteins separated on the gel are usually captured using a charge coupled device (CCD)-camera for image analysis. The gels are analyzed using a program specialized for image analysis of 2-D gels such as PDQuest™ (Bio Rad, USA) or Image Master™ 2D Platinum (GE Healthcare, USA). They are largely automated with minimal manual input needed, thus ensuring objectivity in the analysis. The programs use an algorithm to detect the spots on the gels. The spots are then matched to each other in the different samples and the intensity of the spots are used as a quantitative measure that can be used to differentiate between protein expression in the samples. Upregulated proteins, in the studies included in this thesis, which are found to be of interest can then
be excised from the gel, digested with trypsin and further analyzed with mass spectrometry.

Tandem mass spectrometry (MS/MS)

Mass spectrometry (MS) is a technique by which it is possible to measure the mass/charge ratio (m/z) of ions in a vacuum. This data can be used to calculate molecular masses which in metaproteomics is the molecular masses of peptides from protein samples that have been digested with trypsin. Mass spectrometers have the same basic components, an ion source which converts the analyte into gas phase ions in a vacuum, the ions are accelerated into an electric field towards the mass analyzer which separates the ions according to their m/z ratio which is registered by the detector by recording the impact of the individual ions. Ions can be recorded as intact peptide ions (also called parent ions in tandem MS experiments) or as fragmented ions. Analysis of fragmented ions allow masses of peptide fragments to be determined. These can be used to search against true or theoretical mass data of known proteins in databases, or to derive de novo peptide sequences which can be used for searches against proteins sequences using BLAST or FASTA. Fragmentation is achieved by colliding the peptides with an inert gas in a collision cell by a process called collision-induced dissociation (CID). When using two analyzers in series for separation of ions it is called tandem mass spectrometry (MS/MS) [154]. MS/MS is an important technique for protein and peptide sequencing as well as for identification of post-translational modifications (PTMs). Ionization can be achieved in different ways, the two most common is electron spray ionization (ESI) where analytes are ionized out of solution and dried by passing through a gas, producing gas phase ions and the other is matrix-assisted laser desorption/ionization (MALDI) which sublimates and ionizes the sample out of a dry crystalline matrix using laser pulses [163]. Proteins can be fragmented in different ways, CID is the most commonly used for MS/MS in proteomics research. In CID the peptides are internally heated by repeated collisions with an inert gas, usually argon, leading to internal break along the protein backbone along the C-N bond which results in a number of b- and y-ions. Mass analyzers are of central importance and there are currently five major types being used in proteomics, the quadrupole (Q), ion trap (quadrupole ion trap (QIT), linear ion trap (LIT) or linear trap quadrupole (LTQ)), time of flight (TOF) and Fourier-transform ion cyclotron resonance (FTICR) [164]. The Orbitrap is a new type of mass analyzer which was invented in 1999 by Makarov [165] based on the principle of separating ions in an oscillating electric field [166]. Many hybrids using different coupled mass analyzers have been built for different purposes. The LTQ-Orbitrap is a hybrid mass spectrometer by Thermo Fisher Scientific
and consists of the mass analyzers; a linear trap quadrupole (LTQ) coupled to a C-trap and an Orbitrap and has become a powerful tool in proteomics. The LTQ Orbitrap was used in the experiments presented in this thesis, (figure 12).

Mass spectrometry is used in proteomics for identification of proteins, when this is done by first producing protein peptides and thereafter identification, it is called bottoms-up approach and is usually done by two different approaches; sort-then-break where offline fractionation of the sample is performed followed by direct peptide analysis by “peptide mass fingerprinting”. The other approach is the break-then-sort which has gained huge popularity in the proteomics field and is called “shotgun proteomics”. Protein digestion is performed using a specific enzyme, usually trypsin, which cleaves either after arginine (R) or lysine (K), followed by separation on multidimensional chromatography and finally MS/MS followed by automated database searching [164].

Identification of proteins can be done in a number of ways, one approach is to use the peptide spectra generated from tandem MS and compare them to theoretical spectra from peptides in protein databases using search tools such as SEQUEST [167] or Mascot [168]. It is also possible to use short sequence tags that are extracted and searched against databases allowing for a high error-tolerance. The other is by de novo sequencing where peptide sequences are predicted from the spectra and searched against protein databases.
Advantages of the de novo sequencing is that it is possible to search against peptides whose exact sequence is not identified. However, it is computationally intense and requires high-quality spectra. Protein identification is the overall goal and the peptide sequences using de novo needs to be run against a protein database using search algorithms such as BLAST to find possible homologs [169, 170]. There are several software for de novo sequencing, some of the most popular being Lutefisk [171] and PEAKS [172]. All programs used for identification uses different scoring schemes to find the best peptide match [173]. One important measure that has come about to determine the probability of the match being correct is by calculating the false discovery rate (FDR). One way of calculating the FDR is the target-decoy strategy [174] where the false positives rates are evaluated by searching the peptide spectra against a decoy-target database. Instead of determining for every spectrum which is a correct or incorrect match, a decoy database of the same size as the target database is constructed. For this it is important that the sequences in the decoy and target database overlap as little as possible, this can be accomplished by simply reversing the sequences in the target database. The target and decoy databases are combined into a large database and the peptide spectra are searched against it. The false positive estimations can then be used to derive other measurements that help with evaluation of scoring methods.

In the work conducted in this thesis, (figure 13), paper IV, proteins determined to be upregulated from the 2-D DIGE experiment were digested with trypsin and the LTQ Orbitrap was used for tandem MS analysis followed by de novo sequencing using PEAKS software. The MS analysis was performed by the Core Facility at the Medical Faculty at Linköping University. Identification of proteins using database searches will be further described below.
Construction of library of hypothetical proteins from shotgun metagenome next generation sequencing (NGS) data

To identify proteins and recover the sequence of the protein is crucial for enzyme discovery. Merely identifying the function of a novel protein is not sufficient in enzyme discovery since characterization and, thus, production of the protein is needed to determine the possible usefulness of the enzyme. Thus, to be able to identify the correct full-length sequence it is useful to have access to the translated metagenome of the actual sample as a database to search against. Using a public database, even one as large as NCBI nr, does not make it likely to identify the correct protein since these databases does not include a large fraction the sequences of the of the 99% of microorganisms that
cannot be pure-cultured, constituting the microbial “dark matter”. By creating a database containing the hypothetical proteins of the metagenome of the microbial community from which the protein sample was collected the chances of correctly identifying the protein of interest increases, and the risk of identifying false positives decreases. With the advances in genomic sequencing technology, namely the introduction of next generation sequencing (NGS), sequencing of whole microbial community genomes has been made possible and importantly affordable with the rates for sequencing being low. Cost reductions since 2015 has made the possibilities of large scale sequencing widely available and was a prerequisite for this project, (figure 14).

Figure 14. Graph shows the decrease in cost for sequencing DNA, from 2001 to 2017. Data from National Human Genome Research Institute (NHGRI) Genome Sequencing Program (GSP) (www.genome.gov/27541954/dna-sequencing-costs-data/) April 2018.
NGS is a technology which allows for the sequencing of whole microbial community genomes without using PCR amplification which can cause replicate sequence artifacts leading to overestimation of sequence- and gene-abundances. PCR can also cause single substitutions and deletions leading to frame shifts. There are several different NGS technologies and they differ in their read lengths and the depth of sequencing. The most popular high throughput technology is Illumina which offers shorter reads but deeper sequencing. Longer read lengths are easier to assemble but contain more errors and the depth of sequencing is not as great as for shorter reads. Shorter reads are harder to assemble and there is a higher risk of incorrect assembly. To circumvent this, binning can be performed where the reads are sorted according to taxonomy and assembled in groups to avoid mismatching. Assembly of short reads into longer contigs is needed to be able to identify the complete open reading frame (ORF) of the gene of enzymes of interest. Assembly is either reference-based or de novo. Reference-based assembly uses one or more reference genomes to “map” out new contigs, this approach is used when information is previously known about the genome. De novo assembly is done when nothing is previously known about the genome and is often based on algorithms such as de-Bruijn graphs [175]. The early assembly tools were developed for single genomes and caused many problems during the assembly of metagenomes. Now there are a new generation of assembly tools available to handle previous issues, such as MetaVelvet [176], MetaVelvet-SL [177] and Meta-IDBA [178] which employs binning combined with assembly and k-mer frequencies to detect kinks in the de-Bruijn graph and uses those values as thresholds to decompose the graph into subgraphs. A more efficient assembly is then performed on these subgraphs effectively separating contigs from different species [179, 180]. For database searching using metaproteomic data from MS/MS it is possible to translate the genomic library in all six reading frames into a protein library using nucleotide translation tools such as Transeq. This enables the use of metagenomic data for protein identification.

For analyzing the metagenome and make functional annotation of the genes there are several databases. The online software used in this thesis, in paper III, is the data management repository MG-RAST [181], it utilizes several different databases for taxonomic and functional annotation. The data is displayed as abundance profiles and it provides quality control, gene prediction and functional/taxonomic annotation. To speed up the process, MG-RAST predicts all genes and finds the best homologs in isolated genomes using BLAT (BLAST like alignment tool) excluding all hits with less than 70% identity. All subsequent analysis is then performed on the isolated best hits [179].
For the experiments performed in this thesis (paper III and paper IV) the DNA was extracted from the biogas producing microbial communities kept at metabolic steady state. Sequencing was performed using Illumina technology and assembly was done with MetaVelvet. The libraries of constructed contigs were translated in all six reading frames using Transeq, by which also stop codons are identified. This information was used to identify possible ORFs between methionines and stop codons. Furthermore, the contigs containing many open reading frames were separated into a database of individual hypothetical proteins to enable the use of scoring functions in PEAKS. Identification of novel proteins using this database will be described below.

**Protein identification using database search**

Protein identification is the main goal in proteomics, either for mapping of protein expression and physiological activities by function (i.e. what are the microorganisms doing) or, as in the case of this thesis, for enzyme discovery. Either way it is crucial for proteins to be correctly identified and the final step is to match the de novo peptide sequence tags derived from MS/MS analysis against a database of choice.

There are a number of different databases available of different size and quality. The completeness of the database greatly affects the outcome and a curated database might give more accurate results but contain fewer entries which can affect the identification [182]. Some of the most used databases for identification of proteins from MS/MS analysis are described briefly. The Entrez Protein database from the US National Centre for Biotechnology Information (NCBI) also known as the non-redundant database (NCBInr) is the most complete database but the entries are not as well annotated as more curated databases. There is also UniProt containing both Swiss-Prot and TrEMBL where proteins are manually and automatically annotated, respectively, and trying to keep the redundancy low [183]. Reference sequencing database (RefSeq) is also a non-redundant annotated database from NCBI [170]. BLAST is a search algorithm used to search for sequence similarities against the databases discussed above and is used for identification of proteins. When annotating many sequences simultaneously using BLAST it is possible to automate the process using the Blast2GO software [184].

In the work presented in this thesis PEAKS DB was used for database search of the MS/MS results to identify proteins. The PEAKS DB software utilizes de novo sequencing and provide the de novo peptide sequence tags in its search algorithm [185]. The database used was a translated metagenomic library of hypothetical proteins from the complete shotgun sequencing of the biogas producing microbial communities in
paper III and used for the targeted induction in paper I and paper IV. Hypothetical protein sequences identified by *de novo* peptide sequence tags derived from a protein spot selected as up-regulated, were then functionally annotated using the BLASTp algorithm in Blast2Go searching against the NCBI 

nr database. This in order to identify the function of the identified hypothetical protein of potential industrial interest for cloning and further analysis.
Summary of the papers

Paper I: Applying theories of microbial metabolism for induction of targeted enzyme activity in a methanogenic microbial community at a metabolic steady state

The aim of this work was to establish suitable conditions for further development of a metaproteomic approach for novel enzyme discovery by circumventing pure-culturing and at the same time enabling to target extracellular hydrolytic enzymes of choice in a microbial community.

A mesophilic (38 °C) microbial biogas producing community was used as a model system. The community was fed with a chemically defined medium with readily available nutrients ensuring that the microbial population had no need to secrete hydrolytic enzymes to access nutrients for growth. The chemically defined medium was slightly changed several times over a period of 360 days when a final composition was established which allowed the microbial populations to remain at metabolic steady state. During this time any material from the original sample had been washed out or been consumed and the microbial population was fully adapted to the new conditions of the biogas reactor. Constant monitoring of the biogas production and the methane concentration as well as regular measurements of volatile fatty acids (VFAs), pH, nitrogen content, alkalinity, dry solids (DS) and volatile solids (VS) ensured the viability of the microbial population. Induction of an extracellular hydrolytic enzymatic response
was achieved by first starving the microbial population for 5 days to ensure all readily available metabolites were consumed. Protease production was induced by adding a complex nitrogen source in the form of protease-free bovine serum albumin (BSA) at the same organic load as when fed with defined medium. Protease activity increased on day 2 after addition of BSA and was monitored during the experiment. BSA was added three more times after the previous BSA was consumed which was observed from a stagnation of biogas production. The protease activity did not decrease between additions but rather increased further with each addition of BSA, indicating that proteases were accumulating. The microbial population was also induced with the target of cellulase activity by first starving the reactor for four days followed by addition of a complex carbon source in the form of cellulose (filter paper) and added at the same organic load as when fed with defined medium. Cellulase activity had a longer lag phase than the proteolytic response to BSA, 4 days, after which the cellulase activity kept rising until day 6 when activity started to decline. When the biogas production had reached the starting value a second addition of cellulose was made. The lag phase in this case was only 1.5 days and the cellulose activity was higher at equivalent timepoints to the first addition of cellulose. Importantly, for both experiments, both activities were monitored, and it was found that no crosstalk in activities occurred. Thus, showing that the two activities could be induced individually and independently of each other.

By using a microbial population and maintaining it at metabolic steady state it is possible to manipulate different individual parameters to investigate the effects on the microbial population. The aim of this study was to construct a model system where it is possible to use a microbial population that has suppressed hydrolytic activity to induce a specific response in extracellular production of hydrolytic enzymes for novel enzyme discovery. This was achieved by removing highly accessible monomeric nutrients and adding complex nutrients in its place to induce a targeted response of proteases to BSA and cellulases to cellulose, respectively.
Paper II: Assessment of sample preparation methods for metaproteomics of extracellular proteins

The aim of paper II was to assess different sample preparation methods that could be used for the extraction of extracellular proteins from anaerobic microbial communities maintained on chemically defined medium. In part because almost all metaproteomics studies had been directed to the intracellular proteome and there are very large differences in collecting the concentrated intracellular metaproteome of pelleted/washed cells and the dilute extracellular metaproteome situated in the extracellular liquid. And in part because extracellular proteins could be expected to have solubility properties different from intracellular proteins.

Sample preparation is a crucial step in the investigation of metaproteomes. It is also highly important that the preparation method is compatible with downstream analysis such as liquid chromatography, 2-D gel electrophoresis and MS/MS analysis to be able to identify the proteins correctly. Since metaproteomes investigated come from a wide range of habitats and organisms there is no standard method for protein preparation, rather, a method need to be empirically found for each sample. There are a number of interfering substances in most samples that has to be removed. For environmental samples such as soil this includes e.g. humic substances. Thus, when investigating an extracellular metaproteome the proteins will be highly diluted in the environment of interfering substances and there is a question if a concentration step is needed/possible.

Proteins are highly diverse in their properties such as charge, size, polarity and solubility so the sample preparation method has to be able to capture all different kinds of proteins which poses a huge challenge. The method also has to be able to capture low abundance proteins and produce low background noise to be able to identify all proteins on a 2-D gel.

Samples were collected from the mesophilic biogas producing microbial community maintained at metabolic steady state as presented in paper I. Cells were pelleted and the extracellular liquid was used for sample preparation. Sample preparation was performed on samples concentrated by lyophilization as well as directly on the extracellular liquid without a concentration step. Eight different extraction/precipitation methods were assessed. Four were excluded after quantification and analysis of 1-D gels based on amount of protein recovered and background on the 1-D gels. Four methods were further assessed using 2-D gel electrophoresis where number of spots detected as well as resolution on the gel and background were important criteria. Two different
solubilization buffers were investigated differing between 9.5 M urea or a mixture of 7 M urea and 2 M thiourea.

For our samples of extracellular proteins from an anaerobic microbial community maintained on a chemically defined medium it was determined that extraction directly from solution was possible and preferable since lyophilization also concentrated impurities and contaminating substances. A preparation method involving a phenol extraction and further precipitation using methanol and ammonium acetate was used for further studies. Solubilization buffer including both urea and thiourea was deemed superior to using only urea.

The most common precipitation methods using acetone and trichloroacetic acid (TCA) failed in being able to separate proteins and contaminants to yield good 2-D gels for downstream analysis. The best 2-D gels with the least amount of background and the most clearly visible spots came from extraction methods using phenol. However, it was clear that no single extraction method was able to capture the complete metaproteome showing how crucial the sample preparation is for proteomic analysis, regardless of methods for downstream analysis. Thus, sample preparation should be carefully evaluated when examining a new sample. A combination of preparation methods might be advisable to be able to investigate the metaproteome on a deeper level.

The study showed that it is in fact possible to extract dilute extracellular proteins from microbial communities for studies of the extracellular metaproteome. However, a prerequisite to obtain the results on extracellular proteins as presented in the paper is most likely that the microbial community is maintained on a chemically defined medium to minimize contaminating substances.
Paper III: Shotgun metagenomic analysis of community structure of co-cultured microbial communities under controlled artificial conditions: Implications and applications

The aim of this work was to verify that the composition of microorganisms was representative of microbial communities in anaerobic digestion. That is, so that the feeding with the chemically defined medium did not create microbial communities of only a limited number of microbial organisms or a very artificial community structure. This would otherwise counteract the purpose of maintaining the community on a chemically defined medium to circumvent pure-culturing with the aim of exploring a full microbial community for novel targeted enzymes.

Biogas producing communities consist of a vast number of different microbes acting in concert to finally produce biogas of CH$_4$ and CO$_2$. Microbes initially break down macronutrients (proteins, lipids and polysaccharides) outside the cells by secreting hydrolytic enzymes into their surroundings. Those nutrients are then transported across the cell wall and used as fuel for growth and energy in a number of processes, resulting in the production of biogas. As a model for anoxic environments and by naturally secreting hydrolytic enzymes biogas producing communities are of interest for enzyme discovery.

By keeping the microbial population at metabolic steady state on a chemically defined medium with nutrients provided as monomers (amino acids, fatty acids and sugars), it creates an environment where the microbes do not need to produce enzymes to hydrolyze substances since the nutrients provided are readily available to them in an enzyme suppressed state. Although this provides a condition from which induction of targeted enzymes can be performed, and the extracellular metaproteome can be collected and analyzed, it could also introduce biases in the microbial community. For example, by that microorganisms that normally bind to the normally particulate macronutrients no longer have that advantage, or that opportunistic microorganisms that do not have certain genes for hydrolytic enzymes gain an advantage.

For this purpose, the taxonomic composition of microorganisms of two complete biogas producing microbial communities maintained on the same chemically defined medium at two different conditions was analyzed, one mesophilic (38 °C) and one thermophilic (55 °C). Samples of the two communities were subjected to shotgun metagenome sequencing and the composition of microorganisms were analyzed by taxonomical annotation in MG-RAST. What was found from this analysis was that the microbial community maintained on chemically defined medium at mesophilic conditions was in
fact very similar to natural microbial communities in anaerobic digesters treating a variety of substrates under a variety of conditions. The microbial community maintained at thermophilic conditions was less diverse than expected but was nevertheless composed of the thermophilic microorganisms expected. Thus, it was found that the differences in community structure between the two experimental communities were very representative of a mesophilic and a thermophilic community, respectively. Thus, both reactors were found to be representative for enzyme discovery of enzymes derived from anaerobic digesters and, further, highly likely to yield different enzymes adapted for the two conditions.
Paper IV: Application of metaproteogenomics for targeted identification of novel proteases in complex microbial communities

The aim of paper IV was to apply a metaproteogenomic approach for discovery of novel extracellular enzymes in complete methanogenic microbial communities. In particular to find substilisin-like proteases that to date have found the widest use in industry, and could have potential use in enhancing biogas production from proteinaceous substrates [186].

Samples were collected from one thermophilic and one mesophilic biogas producing microbial community maintained on chemically defined medium under metabolic steady state with the purpose of inducing proteases specifically. Each sample was divided into two batch bottles and were given readily available nutrients that represented the organic load they were normally fed. The difference being that one sample was lacking a readily available nitrogen source. Instead the microbial community was fed a complex nitrogen source in the form of protease-free bovine serum albumin (BSA). The two samples of each microbial community were incubated anaerobically at the same temperature as the they were kept in the biogas reactors and the gas production and protease activity was monitored throughout the experiment.

Similar to the result in paper I, protease activity increased within a few days. When protease activity was judged to have peaked in the induced sample, larger samples were collected from both the induced sample and the non-induced reference for downstream protein extraction with the phenol/methanol method as assessed in paper II. The combined extracellular protein complement of both samples were analyzed and compared by 2-D DIGE and proteins judged as upregulated were excised. From the two experiments 29 spots from the mesophilic community was excised and 22 from the thermophilic community. Proteins collected from the two gels were in-gel digested with trypsin and analyzed by MS/MS. The de novo peptide sequence tags generated from MS data were searched against our own libraries of hypothetical proteins constructed from the shotgun sequencing data of the same two communities from which the proteins had been induced. Sequences for hypothetical proteins that were identified with at least 2 unique peptides were run against the NCBI nr database using the BLASTp algorithm for protein functional annotation. For the mesophilic community 27 out of 29 protein spots were found to contain peptides with homology against proteases or peptidases and in the thermophilic community the corresponding number was 17 out of 22 samples collected. Giving a hit rate of 93% and 77% for the mesophilic and thermophilic populations, respectively. However, because it was extracellular proteins that were targeted these proteases were screened for signal peptides for export. This resulted in 11 and 6 identified
proteases with a signal peptide and conserved domains for proteases for the mesophilic and the thermophilic community respectively. That is a hit rate of 38% and 27% in terms of up-regulated extracellular proteases. In addition, 9 of the proteases of the mesophilic community were located in positions on the gel that corresponded to the mass of the sequence of the identified hypothetical protein. That is, the size of the of these proteins indicated mature and intact proteases that had not been degraded in the anaerobic digester environment and thereby interesting candidate proteases for cloning, production and characterization.

However, although interesting and well worth further research, none of the identified novel proteases was a subtilisin-like protein. Therefore the generated databases of hypothetical proteins were also exploited to search for subtilisin-like proteins [186]. The sequence for subtilisin Carlsberg and a thermostable homolog (thermitase) were used, and both sequences was searched against both databases. By this 95 different hits with subtilisin-like domains and conserved amino acids were found for both sequences in the mesophilic database and 6 hits in the thermophilic database. Because of the high sequence homology of subtilisin and thermitase most hits were found by both proteins. Identified sequences of these hypothetical proteins were analyzed for signal peptide resulting in 18 likely extracellular subtilisin-like proteases. However, the majority of these were considerably larger than subtilisin or thermitase and were probably part of other multi-domain protease active proteins. Nevertheless, 3 of these had sequence length and mass very close to subtilisin and thermitase. When these proteins, derived from the databases of hypothetical proteins of the two communities, were analyzed by BLASTp it was found that the closest homologs were a protein from Methanosarcina sp 795 (99% sequence identity), and from two subspecies of thermophilic Moorella, Moorella mulderi (96% sequence identity) and Moorella glycerini (96% sequence identity). Thus, providing highly relevant subtilisin-like proteases for cloning and further characterization for identifying proteases that may be used for e.g. enhancing biogas production from proteinaceous substrates.

In this study it was shown that proteins of a targeted and induced activity in a complete microbial community can be identified. Not only could the function of individual proteins be derived, but also the complete sequence for the respective identified full-length protein could be retrieved. In this case for the targeted discovery of novel proteases.
Conclusions

The overall conclusions that could be drawn for this thesis are:

By taking a microbial community and feeding it with a defined medium it is possible to establish a community living in a metabolic steady state that continually consume the readily available nutrient without buildup of metabolites and contaminants. It creates a unique situation where the members of microbial community does not need to produce and secrete hydrolytic enzymes to gain nutrients i.e. hydrolytic enzyme production is suppressed at metabolic steady state.

It is possible to induce a targeted metabolic response from a microbial community kept in the enzyme-suppressed state by starving the microbial community of readily available nutrients and then supplying the nutrient which is in deficit in a complex form and thus forcing the microbial community to distinctively produce hydrolytic enzymes to utilize this energy source. This was shown in paper I and further established in paper IV.

Sample preparation is crucial for further downstream analysis in metaproteomics. Since proteins cannot be amplified like DNA it is important to be able to extract as much and as clean protein as possible from the sample to gain a representative image of the metaproteome. Different sample preparations methods give different results and it is important to find the method most suitable for the sample. It might even be considered to use more than one preparation method to cover complementary parts of the metaproteome. For extracting dilute extracellular proteins from a sample collected from
A microbial community maintained on a chemically defined medium, even with monomeric substrates as nutrients, still represent a rich and relevant source of microorganisms for enzyme discovery. In addition, those microorganisms that dominate does so because of the set conditions, not because of the nutrient composition. That is, microorganisms that are normally dominating under mesophilic and thermophilic conditions, respectively, did so also while maintained on chemically defined media. However, how general this finding is, need to be investigated also on other communities to find out if e.g. a complete set of acidophilic, halophilic etc. communities can be maintained in the same way which was concluded in paper III.

In enzyme discovery in microbial communities of microorganisms that cannot be obtained in pure cultures it is important to be able to identify the proteins and it is crucial to have the full-length sequence of the protein of interest in order to be able to produce it in a host organism. By utilizing metagenomics and next generation sequencing it is possible to sequence practically whole metagenomes and assemble them into libraries of contigs. By translating the contigs into amino acid sequences a library of hypothetical proteins of a microbial community can be obtained. By using results from tandem mass spectrometry and searching against the hypothetical products of the actual metagenome of the same sample, the chances of finding the correct sequence increases. Especially so when studying a community of uncharacterized members.

By utilizing approaches from both metaproteomics and metagenomics i.e. “metaproteogenomics” it is possible to identify not only the function but also the complete sequence of novel enzymes which have been upregulated in response to a targeted induction. This is a new approach for targeted discovery of novel enzymes and can most likely be utilized for other microbial populations where enzymes of interest might be found. The key is to initially have a community in enzyme-suppressed state from which a specified and distinct enzyme activity can be induced to be able to discriminate those enzymes that were up-regulated from the large background of proteins that are constitutively expressed at metabolic steady state.
Acknowledgements

Nu börjar slutet närma sig, boken är färdig. Det har varit en berg-och-dalbana av känslor och nu när äkturen närmar sig sitt slut så kan jag se tillbaka på denna tiden och känna mig stolt över vad som kom ut av det. Jag vill tacka alla som varit med mig på denna resan och som på något sett påverkat denna tiden under mitt liv. Jag speciellt vill tacka:

Min huvudhandledare Martin Karlsson som introducerade mig för detta projektet när jag skulle börja mitt masterarbete och som sen gav mig möjligheten att doktorera. Tack för ditt stöd, entusiasm och nya innovativa idéer som lagt grunden för min forskning.

Min biträdsande handledare Nalle Jonsson för stort visat intresse, din entusiasm för forskningen och intressanta frågor som ger nya infallsvinklar och får en att tänka.

Uno Carlsson som presenterade mig och Martin när jag skulle göra mitt kandidatarbete och jag började då med HCA II och det ledde sedan vidare till andra projekt.

Maria Turkina och alla som jobbar på Core Facility på Medicinska Fakulteten vid Linköpings Universitet som hjälpt till med att köra mass-spektrometri.

Björn Wallner som leder bioinformatikgruppen på IFM/Kemi som hjälpt till med assembly av våra sekvenserade metagenom och sortering så att vi fick fina databaser att söka emot.
Inzymes AB som stod för finansiering under halva min doktorandtids och Tekniska verken AB för samarbete och hjälp med reaktorer och mätningar, Yasna Calderon som hjälpt till med provtagning och matning av reaktorerna.

Alla som är del av BRC och framför allt RP1-gruppen som gett mig ett nya erfarenheter av hur forskning och näringsliv kan samverka för att få fram resultat som är viktiga för samhället och Annika Björn för stöd och din entusiasm och för att du tillsammans med Martin Karlsson driver projektgruppen på ett fantastiskt sätt.

Jutta Speda det blev vi två doktorander i gruppen när jag började och jag är så tacksam att det var just dig som jag fick dela kontor och jobba med. Du har varit mitt stöd, du har peppat mig och framför allt har vi haft det så roligt tillsammans. Det har varit en fanstatisk tid och vi har haft så kul under resan och önskar jag att vi kunnat fortsätta jobba ihop för alltid :).

Anna Odnell, det var väldigt kul den tiden du var del av vår grupp och att jag fick dig som vän, tack för alla trevliga samtalar på och utanför kontoret och roliga konferensresor vi gjort. Ironiskt nog blir vi klara nästan samtidigt nu ändå :) Patrik Nygren för dina fina insikter om saker och alla skratt, diskussioner och pubquiz. Lotta Tolstoy Tegler för trevliga samtalar och för din positiva inställning och stöd, det var så kul att få lära känna er och är glad ni var del av vår grupp. Maria Jonsson som kom in emot slutspurten och redde ut och tog tag i alla experiment som behövdes göras och som jag aldrig hade hunnit med, tack för stöd och trevliga samtalar och pepp under skrivandet. Tomas Mårtensson, Anna Brynås som var del av vår grupp under en tid, tack för trevliga samtalar och diskussioner. Jenny Sahlin min duktiga och ambitiösa ex-jobbare, det var kul att få lära känna dig.

Zuzanna Pietras my latest and final office partner during my PhD studies, you are starting out when I am just finishing, thank you for your sarcastic sense of humor and talks during the writing of this thesis. Good Luck with your own thesis. You will do great!


Susanne Andersson tack för allt stöd och trevliga samtalar och administrativ hjälp när saker krånglat. Blev tomt när du slutade.
Undervisandet har varit en stor del under doktorerandet och det har varit fantastiskt kul, Annica Blissing, Amélie Wallenhammar, Alexander Sandberg för alla roliga timmar på labb på Biokemi 1. Magda Svensson och Helena Herbertsson för stöd och hjälp under undervisning på Biokemi 2 och Biokemi 1 och för allt ert jobb med att skapa bra kurser.


Hong Anh och Simon för alla trevliga stunder och tillfällen vi spenderat ihop. Alla samtal och roliga diskussioner som alltid slutar i politik framåt småtimmarna :) Ser fram emot många fler nyår som firas och resor tillsammans.

Mina fina vänner som jag tyvärr träffar alldeles för sällan men som betyder så mycket Karin I, Leila R, Josefin E, Malin J och Sara W.

Farmor för allt du gjort för mig under min uppväxt och ställt upp och hjälpt till och varit där.

Mina svärföräldrar Annica och Ola, är så tacksam för er och allt ert stöd speciellt under skrivandet av denna boken. Det är alltid så skönt och roligt att komma ner till er och ni är såna underbara farföräldrar till Ludvig.

Min bror Alexander för att du finns där, stöttar och hejar när det behövs. Är så glad att jag har dig och hoppas att vi kan ses snart nu när boken äntligen är klar :).

Mamma och Pappa, för allt ert stöd under hela mitt liv, för att ni alltid finns där när man behöver er och för att ni alltid ställer upp och stöttar. Ni har gett mig förutsättningarna att välja och göra det jag är intresserad av. Älskar Er!

Ludvig, min kärlek till dig är oändlig, och även om du ännu inte förstår hur mycket du betyder i mitt liv så ger du mig sån otrolig glädje och energi. Att mötas av ditt glada leende när man kommer hem är något av det bästa jag vet och gör mig alltid så glad.

Mikael min älskade make, utan dig hade detta aldrig gått. Du har stöttat mig genom det hela, dagpendlat till Stockholm bara för att vi ska kunna bo kvar och jag ska ha nära till
labbet, ställt upp när det kärvar på jobbet och tagit hela vabruari för att jag ska kunna
skriva denna boken. Du är där och lyssnar och peppar när jag kommer hem och det
betyder allt. Jag ser fram emot alla nya äventyr som vi ska starta snart.
References


Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-147841