Methods development for metaproteomics-guided bioprospecting of novel enzymes

Jutta Speda

Department of Physics, Chemistry and Biology
Linköping University, Sweden

Linköping 2017
Cover: 2-D DIGE gel used for the accurate identification of extracellular cellulolytic enzymes secreted by a methanogenic microbial community. Red protein spots indicate proteins upregulated by the addition of cellulose.

During the course of the research underlying this thesis, Jutta Speda was enrolled in Forum Scientium, a multidisciplinary doctoral programme at Linköping University, Sweden.
Look deep into nature,
and then you will understand everything better.

Albert Einstein
Abstract

Industrial biotechnology has been announced by several organizations and governments as a key enabling technology for the enhanced economic growth in a low-carbon and knowledge-based bioeconomy. An important goal to promote an environment friendly and sustainable industrial biotechnology is the discovery of new enzymes.

To date, almost all enzymes used in industry have been discovered by pure culturing of microorganisms, however, it is known that less than 1% of all microorganisms can be obtained in pure cultures. The remaining majority of microorganisms is only viable by close biological interactions provided in microbial communities and is not available for enzyme discovery using the classical pure culture approaches. The investigation of microbial communities, which can be viewed as metaorganisms, has been enabled during the last two decades by refining established methods for the analysis of genes, mRNA or proteins and are called metagenomics, metatranscriptomics and metaproteomics, respectively. To date, these techniques have mostly been used in the field of microbial ecology for the understanding of the composition, function and metabolism of microbial communities but not for the purpose of bioprospecting for novel enzymes. Identification of genes that code for possible enzyme candidates is hindered, due to the fact that 30-40% of the sequenced metagenomes contain genes coding for unidentified proteins. Additionally, the -omics techniques generate large amounts of data that need to be analyzed and the outcome of the analysis does not necessarily lead to the discovery of novel applicable enzymes.

The work presented in this thesis describes the establishment of the necessary conditions for a metaproteomics-based method that allows a straightforward and targeted identification of novel enzymes with desired activity from microbial communities. The approach provides a valuable alternative to the incomplete and inefficient analysis of non-targeting data and laborious workflow, which is typically generated by the established meta-omics techniques. In developing the methods presented in this thesis, microbial communities in constructed environments were established, which allowed for the controlled expression of extracellular hydrolytic enzymes under defined conditions. By combination and modulation of advanced metaproteomics and metagenomics techniques, we were
able to directly identify the enzymes and the corresponding gene sequences of several cellulolytic enzymes as a first example for the feasibility of this approach.
Populärvetenskaplig sammanfattning

Utveckling av en ny metod för att hitta enzymer i mikrobiella samhällen

Vi lever i en tid där utveckling av resurseffektiva och miljövänliga metoder för produktion av energi och kemikalier från förnybara råvaror är viktiga mål. Detta mål att ersätta fossila råvaror med förnybara råvaror utgör grunden för bioekonomin. För att göra dessa processer effektiva är biokatalysatorer nödvändiga för att accelerera de underliggande kemiska reaktionerna. Dessa biokatalysatorer kallas enzymer och finns i varje levande varelse. Redan idag används enzymer i olika industriella biotekniska processer, men för att enzymer ska kunna fungera i en teknisk process måste de ha de nödvändiga egenskaperna för detta. Till exempel att vara stabila och funktionella så länge som möjligt vid höga temperaturer. Om en bioteknisk process är möjlig avgörs därför till stor del av om det finns enzymer med lämpliga egenskaper tillgängliga. För att utveckla nya enzymbaserade industriella processer är det därför avgörande att nya enzymer kan identifieras.

En stor källa till nya enzymer är mikroorganismer eftersom de har förmågan att överleva i nästan alla ekologiska nischer på jorden. Dessa innefattar platser med extrema förhållanden, såsom hög temperatur, extrema pH-värden eller höga saltalter. Studiet av mikroorganismer från extrema miljöer är därför särskilt värdefullt, eftersom de producerar enzymer som kan förväntas vara aktiva och stabila under dessa förhållanden, och därmed också kan användas som biokatalysatorer i industriella processer med liknande förhållanden. De enzymer som idag är kända och används i industriella processer är dock isolerade från bara ett fåtal välkända mikroorganismer som man historiskt lyckats renodla i laboratorier. Med nya molekylärbiologiska metoder har man dock insett att den mikrobiella mångfalden är mycket större än den som representeras av renodlade mikroorganismer. Det har uppskattas att endast c:a 1% av alla mikroorganismer kan renodlas, eftersom de bara kan överleva i symbios med andra mikroorganismer i mikrobiella samhällen. Om renodling skall användas för att identifiera nya enzymer kommer man alltså att missa 99% av källan till nya potentiellt användbara enzymer. Utvecklingen av nya renodlingsberoende metoder tillåter numera studier av den totala genetiska och enzymatiska
Populärwissenschaftliche Zusammenfassung

Ein Verfahren für die Entdeckung neuer Enzyme in mikrobiellen Gemeinschaften für verbesserte, umweltfreundliche industrielle Prozesse


Das Ziel der hier vorliegenden Doktorarbeit ist die Entwicklung einer effizienten und innovativen Meta-omics-basierten Methode um neue, technisch relevante Enzyme in mikrobiellen Gemeinschaften zu identifizieren. Da diese Enzyme neu und vorher unbekannt waren, sind sie sehr wertvoll für die verbesserte Anwendung von umweltfreundlichen, biotechnologischen Prozessen.
List of publications

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

(* These authors contributed equally to the work)

Under review in "Biotechnology for Biofuels", November 2016

(* These authors contributed equally to the work)

In manuscript

In manuscript
**Contribution report**

I  Jutta Speda (JS) designed the cellulase induction experiment and co-authored the manuscript.

II JS designed and carried out the cellulase induction experiment. JS was involved in the experimental work and co-authored the manuscript.

III JS developed the 2-DE protocol, contributed to the experimental design and co-authored the manuscript.

IV JS developed the method for cellulase induction and the experimental design. JS performed all laboratory work and analysis. JS co-authored the manuscript.

V JS developed the method for cellulase induction and the experimental design. JS performed all laboratory work, except the next-generation DNA sequencing. JS contributed to the analysis and co-authored the manuscript.
Conference contributions


Thesis committee

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

Faculty opponent
Professor Antonius Van Maris
School of Biotechnology
Division of Industrial Biotechnology
KTH Royal Institute of Technology, Stockholm, Sweden

Committee board

Professor Sophia Hober
AlbaNova University Center
Division of Protein Technology
KTH Royal Institute of Technology, Stockholm, Sweden

Dr. Lars-Göran Mårtensson
IFM, Department of Chemistry
Division of Biochemistry
Linköping University, Linköping, Sweden

Professor emeritus Bo Svensson
TEMA- Department of Thematic Studies
Environmental Change (TEMAM)
Linköping University, Linköping, Sweden
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-D</td>
<td>one-dimensional</td>
</tr>
<tr>
<td>2-D</td>
<td>two-dimensional</td>
</tr>
<tr>
<td>2-D DIGE</td>
<td>two-dimensional difference gel electrophoresis</td>
</tr>
<tr>
<td>2-DE</td>
<td>two dimensional gel electrophoresis</td>
</tr>
<tr>
<td>4-MUC</td>
<td>4-methylumbelliferyl β-D-celllobioside</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>CAZyme</td>
<td>carbohydrate-active enzyme</td>
</tr>
<tr>
<td>CBM</td>
<td>carbohydrate-binding module</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>CMC</td>
<td>carboxy-methyl cellulose</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiotreitol</td>
</tr>
<tr>
<td>EC</td>
<td>enzyme commission</td>
</tr>
<tr>
<td>ESI</td>
<td>electrospray ionization</td>
</tr>
<tr>
<td>EU</td>
<td>European Union</td>
</tr>
<tr>
<td>FDR</td>
<td>false discovery rate</td>
</tr>
<tr>
<td>GB</td>
<td>gigabyte</td>
</tr>
<tr>
<td>HTS</td>
<td>high throughput screening</td>
</tr>
<tr>
<td>IEF</td>
<td>isoelectric focusing</td>
</tr>
<tr>
<td>IPG</td>
<td>immobilized pH gradient</td>
</tr>
<tr>
<td>LPMO</td>
<td>lytic polysaccharide monooxygenases</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>mRNA</td>
<td>messenger ribonucleic acid</td>
</tr>
<tr>
<td>MW</td>
<td>molecular weight</td>
</tr>
</tbody>
</table>
NGS  next-generation sequencing
NHS  N-hydroxysuccinimide
ORF  open reading frame
PCR  polymerase chain reaction
pI   isoelectric point
PTM  post-translational modification
RFU  relative fluorescence units
rRNA ribosomal ribonucleic acid
SDS PAGE sodium dodecyl sulfate polyacrylamide gel electrophoresis
TCA  trichloroacetic acid
TCEP tris(2-carboxyethyl)phosphine
U    enzyme activity unit
USD  US dollar
Vh   volt × hours
Contents

ABSTRACT .................................................................................................................. V
LIST OF PUBLICATIONS ...................................................................................... XI
CONTRIBUTION REPORT .................................................................................... XII
CONFERENCE CONTRIBUTIONS .......................................................................... XIII
THESIS COMMITTEE ............................................................................................ XIV
ABBREVIATIONS .................................................................................................. XV
PREFACE ................................................................................................................ XIX

1. INTRODUCTION ................................................................................................. 1
  1.1 WHAT IS BIOTECHNOLOGY? ........................................................................... 1
  1.2 INDUSTRIAL BIOTECHNOLOGY .................................................................... 2
  1.3 INDUSTRIAL ENZYMES ................................................................................ 3
    1.3.1 Structure of enzymes ............................................................................. 4
    1.3.2 Classification of enzymes ..................................................................... 5
    1.3.3 Industrial enzyme applicability .............................................................. 6
  1.4 THE SEARCH FOR IMPROVED INDUSTRIAL ENZYMES: BIOPROSPECTING AND
    ENZYME DISCOVERY BY META-OMICS APPROACHES ............................... 8
    1.4.1 Metagenomics ...................................................................................... 9
    1.4.2 Metatranscriptomics ............................................................................ 12
    1.4.3 Metaproteomics .................................................................................. 13
    1.4.4 Metaproteogenomics: Combining Genomics and Proteomics .............. 14
  1.5 CARBOHYDRATE-ACTIVE ENZYMES OF INDUSTRIAL INTEREST ............... 15
  1.6 THE BIOGAS PRODUCING MICROBIAL COMMUNITY .................................. 20
  1.7 METAPROTEOGENOMICS- Targeted Bioprospecting for Novel
    Enzymes ......................................................................................................... 22

2. AIMS OF THE STUDY ........................................................................................... 25

3. METHODOLOGY .................................................................................................. 27
  3.1 THE CONSTRUCTION OF METHANOGENIC MICROBIAL COMMUNITIES AT
    STRICTLY DEFINED CULTURING CONDITIONS .......................................... 27
3.2 Induction of Targeted Enzyme Activity in the Methanogenic Microbial Communities at a Metabolic Steady State ...................... 28

3.3 Enzyme Activity Assays ................................................................. 29
  3.3.1 Protease assay .............................................................. 29
  3.3.2 Cellulase assay ............................................................. 30

3.4 2-D Gel Electrophoresis ............................................................... 31
  3.4.1 Isoelectric focusing - the first dimension ....................................... 31
  3.4.2 SDS PAGE - the second dimension ........................................ 34
  3.4.3 Image analysis ................................................................. 35
  3.4.4 2-D DIGE .............................................................. 36

3.5 Zymography .............................................................................. 38
  3.5.1 In-gel zymography of cellulases using Congo Red activity staining ... 39
  3.5.2 Visualization of β-glucosidases by 4-MUC staining ......................... 39

3.6 Tandem Mass Spectrometry ....................................................... 40
  3.6.1 Database searching in tandem mass spectrometry for protein identification ............................................. 44

4. Summary of the Papers .................................................................. 47

Paper I: Applying Theories of Microbial Metabolism for Induction of Targeted Enzyme Activity in a Methanogenic Microbial Community at a Metabolic Steady State .............................................. 48

Paper II: Enhanced Biomethane Production Rate and Yield from Lignocellulosic Substrate by In Situ Anaerobic Digestion Treatment with Endogenous Cellulolytic Enzymes ............................................. 49

Paper III: Assessment of Sample Preparation Methods for Metaproteomics of Extracellular Proteins ............................................. 50

Paper IV: Metaproteomics-Guided Selection of Targeted Enzymes for Bioprospecting of Mixed Microbial Communities ............................................. 51

Paper V: Metaproteogenomics-Guided Identification of Novel Cellulases by Substrate Induced Dynamic Gene Expression in a Microbial Community ............................................. 52

5. Conclusions ................................................................................. 55

Acknowledgements ........................................................................ 59

References ...................................................................................... 63
Preface

With the work presented in this thesis, I have successfully contributed to the development of an innovative, new method that enables the targeted identification of novel enzymes from microbial communities. The use of this method will be an invaluable tool in discovering novel enzymes that are needed as biocatalysts for the successful shift from fossil oil based industry to a sustainable industrial biotechnology, for example, in the production of biofuels from renewable, plant-based materials. Identifying new enzymes in microbial communities is of special interest to the biotechnological sector, due to the possible gain in production/efficiency and lowered costs. Technical limitations have, until recently, prevented close investigations of microbial communities.

Microorganisms exist in communities at very diverse conditions on earth, they also express enzymes that are functional at different conditions and thus of high value for the application in diverse industrial processes. The systematic investigation of microbial communities started in 1998, and has since then been fueled by technological progresses in the bioanalytical field of the so called –omics techniques. To date, the focus of most scientific studies of microbial communities lies on the investigation of metabolic processes and the characterization of the communities for microbial ecology or medical reasons. Studies that focus on bioprospecting for novel enzymes are scarce and suffer from the non-targeting nature of current methods with low identification numbers and hit rates as well as huge amounts of heterogeneous data that needs to be analyzed.
By development of an innovative method that allows for the highly targeted induction and identification of extracellular hydrolytic enzymes from microbial communities, we were able to identify several novel enzymes at a high hit rate.

In the following sections of this thesis, detailed information will be given about the background of the project, the use and limitations of current bioprospecting approaches, the methodology used in this project, and the important conclusions that can be drawn from the work presented. The results are explained in detail in the enclosed scientific papers and also summarized to give a compendious overview of this project.

With my scientific background as an engineer in applied biotechnology, I found it very interesting and motivating to be part of the Molecular Biotechnology group at Linköping University that developed this method for metaproteomics-guided bioprospecting to identify novel enzymes for applications in bioindustries. I have been involved in this project from the very start, when the very idea was put to paper by my supervisors Martin Karlsson, Uno Carlsson and Bengt-Harald Jonsson. Back then we had no experience in the practical application of –omics-based techniques and I am thankful that I was involved in establishing those techniques in our laboratory. This project offered me the possibility for the acquisition of knowledge, as well as immense professional skill development and personal growth.

I hope you enjoy your reading

Jutta Speda,
Linköping, December 2016
“Science should be the most fun job on the planet. You get to ask questions about the world around you and go out and seek the answers. Not to have fun doing that is crazy”

Craig Venter
1. Introduction

1.1 What is biotechnology?

Biotechnology is a multidisciplinary discipline that combines natural sciences with engineering sciences for different purposes, e.g. medical, agricultural or industrial purposes. Natural sciences can be biology, microbiology, chemistry, biochemistry, molecular biology, genetics and bioinformatics, and the engineering sciences such as chemical and process engineering. The United Nation Environment Programme defined biotechnology in their Convention on Biological Diversity as “any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products or processes for specific use” Secretariat of the Convention on Biological Diversity [1].

From a historical perspective, biotechnology has been used since ancient times to produce beverages such as beer and wine; food such as cheese and bread; and seasoning such as soy sauce and vinegar. Different microorganisms were used for those fermentations, but the complex biological processes were not understood and were often associated with mystic concepts [2]. The advancement towards modern biotechnology involved a more detailed, rational understanding of the world and the emergence of science in the 19th century. In 1897, Buchner discovered that alcoholic fermentation does not require the presence of living yeast cells, if catalysts, known as enzymes, are present in the reaction solution [3]. Since then, revolutionary discoveries have been made that allow us to understand
and change biochemical processes of life on a molecular level. Very important examples that illustrate the advances and the deeper understanding are the discovery of penicillin by Flemming in 1928, the discovery of the structure of DNA by Watson and Crick in 1953, the concept of protein synthesis in 1961 by Monod and Jacob, the discovery of bacterial restriction enzymes in 1962 by Arber, and the development of the molecular cloning by Berg, Cohen and Boyer in 1973 [4, 5].

Today, a wide range of different products and applications can be obtained using biotechnological methods and a classification of the different branches using colors is common [6]. This classification is occasionally described as the “rainbow” or “flag” of biotechnology, where each sector is assigned to a certain color. Details are summarized in table 1. It should be noted, that fields may overlap and a distinct differentiation is often not possible.

<table>
<thead>
<tr>
<th>Assigned color</th>
<th>Biotechnological Application Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>Medicine, Pharmacy</td>
</tr>
<tr>
<td>Green</td>
<td>Agriculture, Plants</td>
</tr>
<tr>
<td>Blue</td>
<td>Marine Resources, Aquaculture</td>
</tr>
<tr>
<td>Gray</td>
<td>Waste, Bioremediation</td>
</tr>
<tr>
<td>White</td>
<td>Industry</td>
</tr>
</tbody>
</table>

### 1.2 Industrial Biotechnology

Industrial Biotechnology, also known as White Biotechnology, is the sector of biotechnology that includes the sustainable industrial production of chemicals, materials or energy by using microorganisms or their enzymes.

A very important aspect of industrial biotechnology is the use of renewable resources for the sustainable production of different goods, especially when compared to the fossil oil based petrochemical industry [7]. Currently, a variety of different goods and fine chemicals are produced in industrial biotechnology, some examples are: bioenergy such as bioethanol or biogas, bulk chemicals, paper products, detergents, food and biomaterials [8].

The progress of industrial biotechnology is commonly regarded to have a positive impact on our society, the environment and the economy [9] and has been announced in 2009 as one of six key enabling technologies for enhanced
economic growth in a communication by the European Commission [10]. The importance of industrial biotechnology for economic growth and the society has later been emphasized in a press release in 2012 by the European Commission with the strategic goals to “keep pace with the EU’s main international competitors, restore growth in Europe and create jobs in industry, at the same time addressing today’s burning societal challenges” [11].

Two major fields of industrial biotechnology are the production of large-scale bulk chemicals and bioenergy [12, 13]. The importance of these two sectors became even more evident due to the oil crisis in the 1970’s and the understanding that the world’s oil stock is limited [14]. This initiated a shift from chemical industry and catalysis to industrial biotechnology using renewable materials, e.g. plant-based lignocellulosic feedstocks [15]. It has been estimated that the plant biomass, also known as lignocellulosic biomass, that is annually produced on earth corresponds to 60 billion tons of organically bound carbon and thus represents a huge renewable stock for the sustainable production of bioenergy [16].

The shift from petrochemical industry to industrial biotechnology involved the development of a biofuel market as well as the construction of biorefineries and has since then been in the center of political attention [17]. Two main advantages of the implementation of the industrial biotechnology driven energy sector are the use of renewable materials to produce biofuels, and the reduction of greenhouse gas emission [18]. The burning of fossil fuels increases the net amount in greenhouse gases in the atmosphere, since fossil fuel is derived from lignocellulosic material that grew millions of years ago. In contrast, plant materials used in biofuels have a short growth cycle that causes less net greenhouse gas emission, since approximately the same amount of gases is released as is consumed during growth [19]. Thus, industrial biotechnology offers a sustainable production of bioenergy and chemicals from renewable biomass sources going along with the worldwide political demand e.g. first stated in the Kyoto Protocol and later in the Paris Protocol of the United Nation Climate Change Conference in 2015 of a reduction of CO$_2$ emission [20, 21].

1.3 Industrial Enzymes
Enzymes are proteins that catalyze biochemical reactions. As catalysts, enzymes lower the reaction energy that is required for a biochemical reaction to take place and thus increase the rate of the reaction. Biocatalysts are present in all living organisms and without enzymes, life on earth would not be possible since many
reactions occur too slowly if they are not catalyzed. Enzymes do not only speed up biochemical reactions, they also function as metabolic regulators, and most biochemical reactions are not possible in absence of the enzyme that specifically catalyzes the reaction.

From a historical perspective, enzymes have been used, although unknowingly, since ancient times to produce foods and goods through fermentation, e.g. bread, beer, cheese, wine and for leather bating. In the last century, enzymology evolved as a branch of biochemistry, and the study and understanding of enzymes led to an increased use in industrial settings. Microorganisms are the main source for industrial enzymes and through technical progress in modern biotechnology the development of stable and useful industrial biocatalysts has consecutively progressed [22]. The advantages of using enzymes as an alternative to classical chemical production methods are especially of environmental and economical nature. Enzymes work under mild reaction conditions, e.g. at lower temperature or in environment friendly aqueous settings instead of carbon-based organic solvents, thus lowering energy costs and reducing the amount of potentially hazardous waste and emissions due to their biological degradability as biomolecules [23]. Additionally, enzymes are highly selective in the reaction they catalyze and reduce the number of unwanted by-products thus simplifying production processes [24]. Since enzymes derive from renewable biological resources, their use is a sustainable alternative to classical chemical catalysts and petrochemistry based processes [25]. Today, industrial enzymes are used in new technical applications in laundry detergents, the pulp and paper, textile, and biofuel industries as well as in food processing, the animal feed industry, organic synthesis and the cosmetics industry [26]. The global enzyme market has been estimated to rise to a current value of 7.1 billion USD in 2018 [27], thus clearly showing the importance and rising demand for industrial enzymes.

### 1.3.1 Structure of enzymes

Like all proteins, enzymes consist of amino acids that are connected through peptide bonds and form a long polypeptide chain. The sequence of these amino acids is known as the primary structure. The individual amino acid side chains possess different chemical properties, e.g. charged groups or different polarities that interact with each other and form three-dimensional structures. The secondary structure describes the polypeptide’s local folding structures, e.g. as helices,
sheets or turns, whereas the tertiary structure describes the proteins overall three-dimensional fold. If the protein consists of several individual but interacting subunits, their spatial arrangement is described in the so called quaternary structure [28]. The enzyme’s overall three-dimensional shape is important for its function, since a catalytic reactive site is formed that specifically interacts with the enzyme’s substrate. In order to classify enzymes, it is a common practice to investigate the amino acid sequence and to compare it to the sequence of known proteins, but it is also possible to compare the three-dimensional structures since the functionality is often conserved in so called protein families that possess a similar three-dimensional architecture [29-31].

1.3.2 Classification of enzymes

In the 1950’s, the International Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) was founded by biochemists from the Enzyme Commission (EC), and was given the task to collect and categorize enzyme functions and designed a categorization system, which is still used today. The system uses so called EC-numbers that grade the enzymes accordingly to their chemical function and substrate into six different classes with further specialization into subcategories [32, 33]. Table 2 shows a scheme of the categorization using the first number to sort enzymes by their catalyzed reaction [34].

Table 2. Enzyme Nomenclature by Enzyme Commission (EC) numbers. The table shows the first main EC numbers, the affiliated name and the catalyzed chemical reaction.

<table>
<thead>
<tr>
<th>First EC number</th>
<th>Name</th>
<th>Reaction catalyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oxidoreductases</td>
<td>Oxidations/Reductions</td>
</tr>
<tr>
<td>2</td>
<td>Transferases</td>
<td>Transfer of functional groups</td>
</tr>
<tr>
<td>3</td>
<td>Hydrolases</td>
<td>Cleavage of specific chemical bonds by addition of water (hydrolysis)</td>
</tr>
<tr>
<td>4</td>
<td>Lyases</td>
<td>Cleavage of specific chemical bonds by means other than hydrolysis</td>
</tr>
<tr>
<td>5</td>
<td>Isomerases</td>
<td>Intramolecular stereochemical changes</td>
</tr>
<tr>
<td>6</td>
<td>Ligases</td>
<td>Assembly of two molecules</td>
</tr>
</tbody>
</table>

Table adapted from http://www.chem.qmul.ac.uk/iubmb/enzyme/index.html
For industrial applications, the majority of the enzymes used belong to the EC class 3 and catalyze hydrolytic reactions [22]. In 2002, 70% of all sold industrial enzymes in Europe were hydrolases such as carbohydrases, proteases and lipases. The bulk of those hydrolytic enzymes are proteases that are used in detergent industry to remove soil at energetically favorable low washing temperatures, as well as cellulases, amylases and lipases; all of the named enzymes are also used in the pulp and paper, textile and biofuel industries [35].

1.3.3 Industrial enzyme applicability

The key objective of any industrial process is to obtain a high rate of transformation from substrate into product in the least possible amount of time. This is achieved by using a catalyst that speeds up the reaction. If biocatalysts, i.e. enzymes applied for industrial processes, are used, the efficiency of catalysis is reflected in a high turnover number and thus the broad applicability of the enzyme [36]. Factors that influence the enzyme’s applicability are the physicochemical properties including stability, activity, specificity and efficiency and are unique for each industrial process and each enzyme [37]. Thus, the enzyme’s applicability is not only determined by one single characteristic, but all factors that need to be considered for the applicability in industrial processes. The factors are further explained in figure 1.

The usefulness of an enzyme in an industrial process is contingent on that the enzyme remains active and stable for a long time under the prevalent reaction conditions. Poor stability negatively affects the economics of the catalyzed process and results in longer process operation times, reduced product yields and require increased quantities of enzymes [36].

On the molecular level, high physicochemical stability results in an ideal catalytic activity due to an enhanced intrinsic molecular stability and the maintained functional three-dimensional shape of the protein [36, 38]. The investigation of thermostable enzymes has shown increased hydrogen bonding, structural features such as a high \(\beta\)-strand content, helix stabilization by ion pairs and a high amount of mainly charged amino acids instead of uncharged, polar residues on the surface of the protein compared to, for example, enzymes found in humans [34].
If the amino acid sequence of a good enzyme candidate is known, it is possible to alter this sequence by protein engineering techniques, such as site-specific mutagenesis, protein evolution and rational protein design to further improve the applicability [39, 40]. In the past years, it has been possible to modify chemical manufacturing processes to meet the limited stability of natural enzymes, this, however, may lead to compromised processes [36, 41]. The technical progress has made it possible to engineer customized enzymes for industrial processes, instead of designing the process around the limitations of the enzyme [42], but so far appropriate enzyme engineering is still difficult and limited. Thus, alternative enzyme discovery approaches with improved targeting and identification of novel, interesting enzymes are of immense importance [43]. The greatest range of possible ideal enzyme candidates is believed to be dictated by genetic diversity [36], which leads to novel approaches in the fields of enzyme discovery and bioprospecting by exploration of microbial communities and will be explained in detail in the following section.
1.4 The search for improved industrial enzymes: Bioprospecting and Enzyme Discovery by Meta-omics approaches

Enzymes are widely used in industrial settings for the environmentally friendly and sustainable production of different goods and there is an increased demand for enzymes that work well at the various conditions of different industrial processes. Along with the technical progress in DNA and protein technology during the last decades, different strategies have evolved to meet those growing needs.

As explained in the previous section (1.3.3.), one strategy to improve the performance of known enzymes is by molecular biological techniques such as protein engineering through directed evolution, rational design or chemical modifications [39-41, 44].

Another strategy is the field of bioprospecting, i.e. searching for novel enzymes in various environments that contain microorganisms of high natural diversity. Microorganisms are the most abundant and diverse living organisms that can be found on earth [45, 46]. They are evolutionary adapted to almost all possible conditions and are present in habitats with extreme temperatures, pH, salinity, anoxic environments etc. [47]. Microorganisms that live in extreme environments have evolved enzymes and proteins that are functional under these harsh, prevalent conditions and are therefore highly interesting targets for bioprospecting in enzyme discovery projects. In bioprospecting, the focus does not lie on the chemical variation of known enzymes, but on the investigation of enzymes from organisms or microbial communities found in these extreme habitats in order to identify novel enzymes with desired properties [37]. When interesting enzyme candidates are discovered, the two approaches can be combined, so that newly discovered enzymes from extreme habitats or microbial communities can be further modified and improved on a molecular level [40].

A major reason for the search of novel and better enzymes in microbial communities is that microbial communities present a mainly unexplored supply of new enzymes. Most of the enzymes used in industry today have been isolated from pure cultured microorganisms, but it has been realized that less than 1% of microorganisms can be sustained in pure cultures [36, 48]. The remaining majority of all microorganisms are only viable through close biological interactions provided in microbial communities, or so called metaoorganisms [49]. In the last two decades however, the technical advances in genomics, transcriptomics and proteomics for the large-scale analysis of genes, mRNA or proteins have enabled scientists to systematically study these “metaorganisms” of
interdependent microorganisms. Accordingly, the corresponding techniques applied on microbial communities are called metagenomics, metatranscriptomics and metaproteomics, and are generally known as “meta-omics”. In particular, the development of systems biology meta-omics approaches are instrumental for enzyme discovery in microbial communities.

The investigation of culture-independent methods started only two decades ago and has been tremendously influenced and accelerated by the technical advances in the methods used, namely the development of next-generation sequencing (NGS) of DNA for metagenomics and –transcriptomics [50] that led to an increase in investigated habitats due to reduced costs and improved data quality. In addition, immense progress has been made in field of metaproteomics [51], especially with the development of more sensitive instruments for mass spectrometry, and of improved computational methods for the analysis of the complex samples. Furthermore, the progress in bioinformatics and the development of improved tools and platforms for data analysis along with enhanced computer power, have led to an increased research activity in the field of meta-omics [52, 53].

However, to date, the focus of many scientific studies in the meta-omics field lies on the investigation of metabolic processes and the characterization of microbial communities for microbial ecology or for medical reasons. Studies that focus on bioprospecting for novel enzymes are scarce and suffer from low identification numbers and hit rates. These problems are caused by the complexity and huge amount of data that is generated and the lack efficient analysis methods, which allow for the identification of interesting enzymes against the enormous background of other proteins and genes that are present in the microbial community.

The individual –omic techniques, the combination of those and the approach that is the main theme of this thesis will be explained in detail in the following sections.

1.4.1 Metagenomics

The field of metagenomics refers to the study of the total genetic potential of microbial communities and is also known as ecogenomics, environmental genomics or community genomics.
The genomic profiling of microbial communities circumvents the limitations of studying individual microorganisms and has been termed “metagenomics” by Handelsman et al. in 1998. Metagenomics experiments involve the extraction, cloning and analysis (“screening”) of the entire genetic complement of the microorganisms that are present in a habitat and gives access to the genetic information of unculturable microorganisms [50]. It should be noted that the term “unculturable” typically refers to the inability to obtain pure cultures of single organisms [51]. The term is rather misleading, since many “unculturable” microorganisms can in fact be cultured if they are co-cultured with other microorganisms. The efficient analysis of this entire genetic complement of microbial communities is complicated by the vast complexity and the amount of data that is generated. For the purpose of bioprospecting for novel enzymes, sequence-based metagenomic screening suffers from the lack and quality of available genetic information, and activity-based metagenomics from difficulties in expression of genes against a heterologous background. These problems will be explained in the following text. For the analysis of the generated genomic information from microbial communities, there are two general types of screening methods used: the sequence-based or so called homology-based screening, and the function-based or activity-based screening.

The first screening approach is called homology-based screening or similarity search, since the search is conducted against a reference database. A typical result from a metagenomic NGS experiment is a large data file that contains the assembled sequences of a vast number of DNA pieces of varying length, so called contiguous sequences or contigs. This information needs to be analyzed and it is common to annotate the function of the included genes by comparing the DNA fragments to already known genes. The result depends on the quality and the information that is included in the reference database used. The gene in the metagenome can only be annotated if a similar gene or protein is present and correctly annotated in the reference database. However, around 30-40% of the genes analyzed by metagenomics projects code for proteins that have no assigned function when investigated through sequence homology-based screening and comparison to known proteins [52-54], or do not show enough similarity to already known proteins and enzymes present in the database [55]. These proteins are thus not targeted for by enzyme discovery using these homology-based approaches. The homology-based screening analysis also includes the prediction of genes by identification of their gene coding open reading frame (ORF) in the
DNA fragments. This is usually done by using ab initio gene prediction algorithms, but may be difficult since the genetic information that codes for the start- or the stop region of the gene may be misleading and inconclusive [56, 57]. Consequently, unrecognized or incorrectly predicted genes are not available for homology-based comparisons and will, consequently, be omitted when bioprospecting of novel enzymes is performed. The homology-based approach requires the investigation of the complete metagenomics data set in order to identify interesting enzyme candidates. Therefore, a large search space is created and the results include potentially all genes that are present in the investigated microbial community, however, no information is given about which genes are actually expressed under certain conditions and stimuli. For example, if a metagenome is screened for all proteases present in the microbial community, several hundreds of proteases may be found because they are present in the genomes of the many microorganisms present in the microbial community. Thus, the result contains many different genes with the same function. For bioprospecting purposes however, it may be interesting to e.g. target only extracellular proteases, since these are stable outside the cell and possess interesting biochemical properties for industrial applications. In the experimental hit list, such information is not given and a subsequent characterization of all present enzyme candidates is highly untargeted, expensive and labor-intense.

The second approach in metagenomics for enzyme discovery is the functional or activity-based screening analysis. The activity-based screening analysis involves the expression of genes that are cloned from a metagenomic library and a subsequent scan for a certain enzyme activity [58-60]. Using this methodology, the enzyme’s activity that has not been predicted by homology-based comparisons may be discovered, but it involves the analysis of the protein products of all genes that are found in the sample and typically involves screens by optical detection of the enzymatic activity [52]. Thus, all genes need to be expressed in clones and the approach, even if optimized for robotic high-throughput screening (HTS), is extremely time and cost consuming and technically complicated. These technical difficulties complicate efficient bioprospecting, since for a complete analysis of the metagenome (i) only complete genes may be cloned, (ii) the genes to be cloned need to be correctly inserted into expression vectors, (iii) soluble expression requires host microorganisms that are able to express all foreign genes, which is currently not possible, (iv) the process of enzyme activity screening is limited to
the availability of HTS-compatible assays and (v) it is only possible to screen for one enzyme activity at the time.

In conclusion, it has been shown that both metagenomic methods (sequence homology and activity-based) for enzyme discovery produce on average one active clone in approximately every 1200 clones investigated, which results in a hit rate of 0.083% [54]. This low number illustrates that metagenomics experiments result in too much, irrelevant and incomplete information for the purpose of targeted enzyme discovery. However, the value of the metagenome can be seen in the static information that is provided, which stands for a complete genetic pool and thus the genetic potential of a microbial community. It can therefore be concluded that metagenomics approaches are well suited for e.g. taxonomic studies of microbial communities that investigate the composition of present microbes, but less so for targeted enzyme discovery.

1.4.2 Metatranscriptomics

The area of metatranscriptomics addresses the analysis of all mRNA extracted from a microbial community and contains information about genes that are expressed at the time of sampling under certain conditions [61]. Therefore, the information gained is of a more quantitative and dynamic nature [62] and the communities reaction to stimuli can be investigated. Nevertheless, all mRNA that is present under certain conditions needs to be analyzed, including the mRNAs coding for proteins for e.g. cell reparation processes or the maintenance of other cellular processes, which may not be interesting in the search for novel biocatalysts. Another drawback when working with mRNA samples, is the low chemical and biological stability of RNA that complicates the isolation and handling in the laboratory and makes the working with RNA samples difficult. RNA is very prone to the degradation by the enzyme ribonuclease that is ubiquitous and difficult to inactivate in the presence of RNA and it has been shown that some mRNAs have half-times as short as 30 seconds [63]. Also, the single-stranded nature of mRNA makes the molecules vulnerable to shearing forces. Furthermore, prokaryotic mRNAs generally lack the poly(A)-tails as a structural element that makes the isolation of their eukaryotic counterparts very selective and straightforward [64], in addition are mRNA molecules far less abundant than rRNAs in total RNA extracts and the mRNA signal is often overshadowed by the strong rRNA background signal [62]. Finally, as also observed in metagenomics experiments, the generation of NGS data leads to huge
data files and an efficient analysis needs effective analysis algorithms, much computational power and also ideally complete protein reference databases for correct gene annotations [65].

1.4.3 Metaproteomics
The field of metaproteomics gives insights into the proteins that are expressed within a microbial community at a certain time and can be understood as the “large-scale characterization of the entire protein complement of environmental microbiota at a given point in time” [66]. A major advantage of this method is that the proteins that are analyzed stem from genes that are actually expressed under certain environmental conditions and thus, makes it possible to analyze the dynamic response of the microbial community to defined conditions, a fact that is used in bioremediation experiments for the investigation of pollutant effects and possible neutralization of those [67, 68]. A drawback of metaproteomics is the lack of a general sample preparation method that functions well for all samples, meaning that for each targeted environment, an individual sample preparation method needs to be established. This problem is connected to the fact that unlike DNA, the amount of proteins in the sample cannot be increased or amplified by PCR, so the extracted amount of protein from the sample must be sufficient for subsequent analysis and at the same time ideally represents the present protein pool in the microbial community [69]. Thus, a feasible sample preparation method is essential for the success of metaproteomics.

The metaproteomics workflow consists of the sample preparation step and the subsequent analysis of the proteins that are contained in that sample. Nowadays, it is common routine to separate the proteins by gel electrophoresis or chromatographic methods and subsequently analyze them by mass spectrometry. Peptides that are detected by mass spectrometry are usually compared to known peptides in public databases for annotation of the protein and the protein function by amino acid sequence homology-based approaches. By using public translated gene or protein databases, it is very likely to match the peptides to protein homologs, which means that their amino acid sequence in certain conserved positions is similar to a known protein. However, their actual complete genetic code given by the amino acid sequence may still differ in comparison to the investigated sample. If the exact complementary gene or protein is not present in the database used, only the function of the investigated protein can be identified by similarity to the known protein. This problem can be expected for the majority
of proteins derived from microbial communities of unknown community structure. The lack of correct sequence information may not be a problem for microbial ecology experiments, which mainly investigate which kind of proteins are present under certain conditions. However, for the search of novel biocatalysts, the correct amino acid sequence information is crucial. Only if the correct amino acid sequence is known, is it possible to clone and produce the discovered enzymes with the correct properties. The problem with incomplete protein or gene reference databases and missing sequence information for microbial communities is a reoccurring drawback for all individual meta-omics approaches.

The application of metaproteomics for bioprospecting is further complicated by the immense complexity of the investigated protein samples. For example, of the 4191 predicted genes present in the complete genome of the well characterized model organism *Escherichia coli*, 2800, or approx. $\frac{2}{3}$ of all genes, are estimated to be expressed at any time [70]. If these numbers are applied to a microbial community of e.g. 100 different species, each with an average genome of 4200 genes, there would possibly be 230,000 proteins present at any time. If the change in protein expression between two different conditions is investigated, the expression of any or all of potentially 420,000 proteins could change. Thus, to make any sense of the dynamic change in a protein expression pattern in response to an external stimulus, the experiment must be under very close control so that the response is not due to too many influences. Furthermore, a very sensitive detection method is required to recognize the dynamic change in the protein expression pattern and to differentiate these upregulated proteins from other present proteins and proteins that originate from the surrounding microbial habitat.

### 1.4.4 Metaproteogenomics: Combining Genomics and Proteomics

When all the drawbacks of the individual –omics approaches are considered, it becomes evident that a combination of the different techniques would be beneficial and this combination is known as the integrated –omics approach.

The combination of proteomics and genomics approaches is called proteogenomics and was introduced by Jaffe *et al.* in 2004 [71]. If microbial culture-independent samples are investigated the approach is called metaproteogenomics. In proteogenomic MS/MS experiments, customized protein sequence databases generated from the genomes of the actual samples are used in database searching, instead of public protein databases [72]. To create customized protein databases, all six reading-frames of the genome that is studied are
translated into amino acid sequences and used as a database for peptide matching in the proteomic mass spectrometry experiment [73, 74]. This way, the peptides identified by the tandem mass spectrometry experiment can be directly linked to the gene sequence that codes for these peptides, since both genes and peptides have been obtained from the same sample [75]. The matched peptides enable the localization of exactly those genes that are expressed to proteins, thus confirming coding regions in the genomes and help to refine gene prediction and gene annotation models [76, 77]. Furthermore, proteins can be discovered that may be missed using the metagenomics homology-matching approach and the *ab initio* approach by using gene prediction algorithms as explained above. The advances in computational biology coupled with the increased availability of genomic data has elevated the importance of proteogenomics as a tool for functional gene annotation and it is believed that proteogenomics will become even more important in the future [78, 79].

An impressive example of the potential of this method is a large-scale study by Venter et al. who performed a proteogenomics analysis on 46 bacteria and archaea [80]. The study found evidence for 682 novel proteins, which clearly shows the deficiencies in traditional gene annotation approaches and illustrates the advantages of using the proteogenomics-based approach.

For the purpose of bioprospecting for novel enzymes, the metaproteogenomics approach is very promising, since it leads to the exact gene sequence of the investigated protein samples and thus allows for the production and characterization of the novel biocatalysts. However, a literature search on the term “metaproteogenomic*” during the writing of this thesis in November 2016 revealed that only 14 studies have been published in scientific journals. In these studies, the focus was on the general technical approach of this method [81, 82], microbial ecology and the study of non-model organisms [75], plant biology [83, 84], bioremediation [85] or medical applications [86], and no studies for enzyme discovery were listed. For the purpose of targeted bioprospecting for novel enzymes of industrial interest, the metaproteogenomics approach has not been used yet, which emphasizes the importance and novelty of the method that has been developed in this project and will be further explained in section 1.7.

### 1.5 Carbohydrate-active enzymes of industrial interest

Carbohydrate-active enzymes are important targets in enzyme discovery and since we have focused on the discovery of those enzymes in the enclosed papers IV and...
V, a short description of their value and mode of action will be given in this chapter.

Plant biomass, also called lignocellulosic biomass, is regarded as the most abundant biopolymer on earth and as an important source of fermentable sugars for industrial uses [87, 88]. As mentioned earlier, it has been estimated that the lignocellulosic biomass that is produced annually on earth corresponds to 60 billion tons of organically bound carbon and thus represents a huge stock for the sustainable production of bioenergy [16]. Therefore, the production of bioenergy from lignocellulose has been in the focus of politics and research for the development of a sustainable society and economic growth. The efficient production of bioenergy and biochemicals requires the availability of carbohydrate-active enzymes, which are able to hydrolyze lignocellulosic biomass to fermentable sugars.

Lignocellulosic biomass is composed, mainly, of three different polymers: cellulose, hemicellulose and lignin. The specific composition varies in different plant species, but all plant materials have in common that they are highly recalcitrant to the hydrolytic breakdown by enzymes. Due to this recalcitrant property and in order to enhance the degradation of cellulosic material, it is common to pre-treat the material by either mechanical, thermal or chemical means [89, 90]. The biomass structure is disrupted upon pretreatment and the accessibility for the enzymatic degradation is thus enhanced [91].

Cellulose is a carbohydrate polymer that consists of β-1,4-glycosidic linked D-glucose molecules. The long, unbranched polymer chains can be organized in different ways: in highly organized so called crystalline regions and in less tightly packed, so called amorphous regions. The crystalline regions contain polymer chains that align to neighboring strands by intermolecular hydrogen-bonds and Van-der-Waals interactions. They are arranged in a tight, stack-wise manner, thus forming long, crystalline microfibrils, which are especially recalcitrant to enzymatic hydrolysis. The amorphous regions are less tightly packed and are more accessible for enzymatic degradation.

Hemicellulose exists of different carbohydrates that form, in stark contrast to cellulose, a branched polymer chain. Depending on the plant species, hemicellulose contains different amounts of C₅ sugars, such as xylose, arabinose and rhamnose, and C₆ saccharides as glucose, galactose and mannose. A common trait is the β-1,4-glycosidic and occasionally β-1,3-glycosidic linked backbone that consists of the different hexoses. Compared to cellulose, the polymer chains
are branched and in general shorter and therefore of lower molecular weight. The branched structure also hinders the polymer chains from forming tight crystalline packing, thus making hemicellulose less recalcitrant to degradation than cellulose [92].

Lignin is a polymer found in plant cell walls and consists of cross-linked phenolic residues, namely phenylpropanes coniferyl, sinapyl alcohol and p-coumaryl. Lignin can be found in all plants and has a high content in woody lignocellulosic biomass, especially in the primary cell wall as bark. It is important for the structural support and high resistance to microbial degradation of the plant.

The polymers interact with each other to form lignocellulose and the specific composition differs depending on the plant species. Examples are illustrated in figure 2.

![Figure 2. Examples for the different percentage composition of biopolymers in lignocellulosic plant biomass and substrate for the production of biogas. The varying percentage content in respect to cellulose, hemicellulose and lignin is dependent on the plant species and an important parameter for the recalcitrance in a digestion process. Values adapted from Sawatdeenarunat et al. [93] and Schroyen et al. [94]](image)

Different bacteria and fungi, but also some animals have evolved enzymatic systems for the degradation of lignocellulosic biomass [95, 96]. Due to the complex and recalcitrant nature of the substrate, a synergistic interaction of different carbohydrate-active enzymes (CAZYmes) is needed and it is believed, that the complex process is yet not fully understood. In the classical view of the
decomposition of the insoluble lignocellulosic biomass, it is generally believed that most aerobic microorganisms secrete free cellulolytic enzymes, whereas the majority of anaerobic microorganisms produce large multimodal cell-bound complexes called cellulosomes, but exceptions from this view have been recognized [97, 98].

To date, the interaction of at least four different enzyme families is known to contribute to the degradation of cellulose: endoglucanases, exoglucanases, β-glucosidases and lytic polysaccharide monooxygenases (LPMOs). Endoglucanases bind randomly to the amorphous regions of cellulose and hydrolyze the substrate in a processive manner, which means that the enzyme binds to the substrate and catalyzes repeated reactions without releasing itself in between each successive reaction. The endoglucanase creates “loose” ends for the binding of exoglucanases, which bind to the resulting reducing or non-reducing ends of the cellulose polymer. Upon hydrolysis, exoglucanases release the disaccharide cellobiose that acts as the substrate for β-glucosidases, which degrade the disaccharide to glucose. The involvement of LPMOs in the hydrolytic degradation of other recalcitrant polysaccharides was discovered only a few years ago [99]. They act by producing oxidative species that react with the crystalline regions of cellulose, thus creating binding sites for endoglucanases by loosening the tight, crystalline structure by oxidative cleavage [100, 101]. Since lignocellulosic biomass is an insoluble substrate and enzymes need to be in close spatial proximity to perform the catalytic action, most CAZymes possess a substrate binding site called carbohydrate-binding module (CBM). Thus, many cellulases are built in a modular structure, meaning that they possess one or several catalytic modules as well as one or several binding modules. The substrate specificity differs within the enzyme families, and cellulases may also possess different catalytic activities and binding modules. In some anaerobic bacteria and fungi, the cellulolytic enzymes are arranged in large multicomponent high molecular-weight complexes that bind the bacterial cell to the substrate. These complexes are called cellulosomes [102]. In cellulosomes, several different catalytic and non-catalytic modules are arranged in close spatial proximity leading to a high concentration of catalytic sites, and the specific composition of the modules is adjusted to the available substrate [103]. The synergistic effect among the cellulosomal enzymes themselves as well as the cell is proposed to be beneficial for the cellulosome possessing microorganisms, since enzyme inhibitory products as e.g. cellobiose and glucose are readily consumed by the
close-by bacterial cell [104]. This way, the bacteria that carry the cellulosome benefit from the spatially closeness to the nutrients that are made available and the enzymes are not hindered in their catalytic activity by product-inhibition.

Carbohydrate-active enzymes are grouped into families by amino acid sequence similarity and the latest information e.g. on the sequence-based family classification, three-dimensional structure and sequenced genomes is available online from the Carbohydrate-Active Enzymes database (CAZy database, http://www.cazy.org) [105].

In 2009, cellulases were regarded as the third largest global industrial enzyme traded by dollar volume and are expected to become the largest volume in the future [106], which emphasizes the need and economic potential of novel effective enzyme systems. Industrial cellulases are commercially available in form of mixtures that contain different ratios of carbohydrate-active enzymes and some major global enzyme producing companies include Novozymes (Denmark), Genencore (USA), Danisco (Denmark) and DSM Innovations (The Netherlands), among others [107]. The enzyme cocktails are mostly produced by the aerobic fungi \textit{Trichoderma reesei}, which has undergone immense strain optimization during the last decades and is able to secrete as much as 100 g of crude cellulases per liter of culture broth [107]. A major problem with the use of these enzymes, aside from the high enzyme costs, is the reduction of catalytic activity of the enzymes in the fermenters, due to the conditions in the fermentation mixture. Another problem is the degradation of the enzymes by proteases that are part of the biological system the cellulases are produced and used in [108], this is especially true in biogas production. Thus, for example, to enhancing the biogas production process, it may be meaningful to search for stable cellulases in the same system or technical process they will be used in. The biogas production process is a synergistic anaerobic digestion of organic matter by a community of anaerobic microorganisms. The reducing conditions as well as the proteases produced by members of the microbial community have a negative effect on the activity of endogenously added “foreign” cellulases. It is therefore imperative to search for active cellulases in methanogenic microbial communities with cellulolytic properties, a concept that has been made possible with the recent technical advances in the field of environmental biology and meta-omics, as mentioned earlier.
1.6 The biogas producing microbial community

The anaerobic degradation of organic biomass to biogas is a multistep process that is accomplished by syntrophic microbial communities found in different habitats in nature. Examples of natural habitats are marine and freshwater sediments, gastrointestinal tracts of animals (e.g. ruminants, termites, woodworms), man-made habitats such as flooded soils (e.g. rice fields), organic waste deposits and anaerobic digesters [109].

Biogas consists mainly of methane and carbon dioxide with a minor content of hydrogen sulfate, nitrogen and ammonia. Biomethane is a valuable biofuel used for heating, electric energy generation and vehicle fuel. It has a number of advantages over other alternate fuels such as bioethanol, biodiesel or hydrogen. For example, the combustion of biomethane results in lower emissions of noxious by-products than the burning of biodiesel [110]. Furthermore, biomethane contains approximately three times as much energy as hydrogen, and additionally, storage and transportation of biomethane are safer and more convenient [109]. Another advantage with biomethane is that it is biologically possible to produce the gas from a variety of different wastes and renewable products of less defined composition, such as slaughter house waste, wastewater treatment plant sludge, lignocellulosic biomass and algae among others. In contrast, the sole substrate for the production of bioethanol are monomeric C\textsubscript{6} sugars. The overall important, environmental and economic aspects of biofuels have been described earlier in this thesis, but it should be emphasized again that an increasing demand for alternative energy strategies is expected in the future. This demand will lead to a growing need for efficient hydrolytic enzymes for improved biogas production.

The production of biogas by anaerobic microorganisms is a multi-step process that requires the close cooperation of several anaerobic bacteria and archaea [111] and is illustrated in figure 3. The end product methane is only produced if all key members of the community are viable and present in the correct ratios. Thus, this balanced process involves microbial dependencies and interactions, where different species are living on the products of other microbes, a phenomenon known as syntrophy or microbial cross-feeding [111, 112]. This means that if one step of the process cannot be performed by certain members of the microbial community, the other members will not be able to survive and the process will come to a halt.
Figure 3. The synthropic multistep-process of anaerobic digestion. In order to produce the biogas methane from complex biopolymers, all four reactions of anaerobic digestion have to successfully be performed by all members of the microbial community. Figure modified from Herbst et al. [113]

The first step of anaerobic digestion is the enzymatic breakdown of organic biopolymers (polysaccharides, fats, proteins) into soluble monomers (monosaccharides, long chain fatty acids, amino acids). This degradation is accomplished by bacteria that secrete hydrolytic enzymes with the required hydrolytic function. If recalcitrant substrates are used, the hydrolysis is regarded as the slowest and therefore rate limiting step of anaerobic digestion [114]. This emphasizes the importance and need for efficient hydrolytic enzymes for improvement of the commercial biogas production processes from recalcitrant substrates. The products from hydrolysis are successively processed in the acidogenesis or fermentation step, resulting in the formation of alcohols, organic acids, long chain fatty acids, CO$_2$, H$_2$ and NH$_4^+$. The formation of acetate from organic acids, alcohols and the reduction of CO$_2$ is accomplished in the third step called acetogenesis performed by hydrogen producing acetogenic bacteria. In the last step, methane is produced from acetate, alternatively from carbon dioxide and hydrogen, by different groups of methanogenic archaea. The methanogenic food
chain is thus dependent of the hydrogen transfer from acetogens to a certain group of methanogens, a process called interspecies hydrogen transfer.

Methanogenic microbial communities thus naturally possess extracellular enzymatic systems that are adapted and capable to hydrolyze biopolymers in the prevailing environment of the methanogenic community. A biochemical identification and production of these enzymes could therefore help to improve biomethane production rates if the enzymes could be produced and added to the process. As described earlier in this thesis, the identification of enzymes from microbial communities by meta-omics has only emerged during the last decades and is a powerful method to gain access to novel enzymes of unknown sequence. This wealth of, most probably, yet unknown hydrolytic enzymes and the availability of meta-omics analysis tools, thus, make methanogenic microbial communities an outstanding interesting target for the identification of novel hydrolytic enzymes.

1.7 Metaproteogenomics- targeted bioprospecting for novel enzymes

Metaproteogenomics is a rather new technique in the integrated –omics field and has not yet been used for targeted enzyme discovery. With the work presented in this thesis, a significant contribution has been made to the scientific field of bioprospecting by development of an innovative metaproteogenomics-guided method that overcomes the limitations of individual meta-omic approaches as described before and allows for a highly targeted and efficient identification of novel genes of biotechnological interest from microbial communities [115].

For this purpose, methanogenic microbial communities were chosen, not only because the industrial importance of anaerobic digestion for biomethane production, but also because of the highly syntrophic properties of the involved microbes [116, 117]. The syntrophic dependence of the microorganisms can be used as a control system for the condition that the community is in, because only if the end product of the anaerobic digestion process methane is produced, all members are viable and functional in the community. Details on the methodology are given in the following section and are presented in the included papers. What is important to note is that a viable microbial community was established, which was maintained at a metabolic steady state, which in turn produced an environment with very low background of extracellular hydrolytic enzyme-activity. By exchanging the substrate during culturing, the community responded measurably to the change by expression of extracellular, hydrolytic enzymes to
specifically break down the substrate added, without interference of contaminating substances and proteins that originate from the original habitat (paper I [118]). This is of outstanding importance, because it is usually very difficult in meta-omics projects to identify relevant enzymes or target the production of specific enzymes against the large background of all proteins that are collected. In addition, by maintaining the microbial community in a constructed environment on a chemically defined medium, the possibility to collect and analyze the extracellular metaproteome, or the metasecretome, was realized (papers III [119])

Figure 4. Metaproteogenomics-guided bioprospecting for novel enzymes. A microbial community from a biogas digester was kept viable at a metabolic steady state. This metabolic steady state allowed for the induction of a targeted hydrolytic enzyme activity in a batch experiment by addition of enzyme activity inducing substances. By comparing the metasecretome of the induced microbial community to an un-induced control sample in a 2-D DIGE experiment, significantly upregulated proteins in the induced sample could be analyzed and selected for a MS/MS identification. By using the translated metagenome of the microbial community as a database for protein identification, the exact gene sequence of the upregulated spots was determined and thus allows for the molecular cloning of the targeted and identified enzymes [115].

The upregulated extracellular proteins in the experiment could be visualized and measured by application of sensitive, advanced gel electrophoresis based proteomics techniques (papers III [119], IV and V). The metagenome of the
microbial community was sequenced, translated into all six reading-frames and used as a database in a following tandem mass spectrometry experiment for the investigation of the upregulated proteins. This way, the method enables the identification of the exact gene sequence of the proteins upregulated due to the altered culturing conditions and allows for the production of these exact enzymes by molecular methods (paper V). The discovered enzymes are most probably involved in the breakdown of the new substrate and need further characterization. Since the approach is highly targeted, the number of enzymes for characterization in rather low and there are typically between 4 to 8 candidates that need to be investigated further, out of 30 to 40 proteins analyzed. In the initial experiments for the identification of novel cellulases, the hit rate for the functional identification of the targeted, upregulated proteins was between 44% and 96%. In the final metaproteogenomic experiment, it was possible to identify the exact gene sequence of four different novel cellulases. This genomic identification resulted in a hit rate of 16.7%, which is notably higher than in established meta-omics approaches that show a hit rate of 0.083% [54]. A flow-scheme of this method is illustrated in figure 4.

Compared to the other omics-approaches, the selection of possible enzyme candidates is furthermore accomplished by upregulation as identified on the electrophoresis gel and not due to a predicted sequence-homology to known proteins. Therefore, the selection also leads to the identification of enzymes that were not earlier known to participate in the degradation of the new substrate (paper IV). Additionally, it is possible to identify enzymes that work complementary and synergistically in the hydrolytic decomposition of the new substrate that allows for the exact composition of so called enzyme cocktails, which are increasingly used in the industrial decomposition of lignocellulosic biomass (paper IV and V).

This metaproteogenomics method allows for the identification of novel extracellular hydrolytic enzymes in microbial communities with a high hit rate, precision and accuracy against a huge background of other, irrelevant proteins, for the purpose of enzyme discovery.
2. Aims of the study

The overall aim of this study was to develop a metaproteomics-based method for bioprospecting of targeted novel extracellular hydrolytic enzymes from microbial communities. In addition, to be useful as an enzyme discovery method for biotechnology, the gene coding for targeted enzymes should also be identified to allow the production of the identified proteins. The steps to reach this objective involved the development and application of several necessary steps as listed below.

(i) The construction of a viable methanogenic model microbial community in a clean, fully defined environment that allows for the control over the expression of extracellular hydrolytic enzymes, so-termed induction, against a clear sample background (paper I)

(ii) The verification that the induced extracellular cellulolytic enzyme system have an effect on biogas production rate and yield from lignocellulosic substrate (paper I and paper II)

(iii) The assessment and verification of efficiency of sample preparation methods for metaproteomics studies by 2-D gel electrophoresis of extracellular proteins (paper III)
The application of the methods above, in combination with 2-D DIGE and tandem mass spectrometry to verify that the methods allowed for the selection and analysis of specifically upregulated and secreted enzymes due to the induction of cellulolytic enzymes (paper IV)

The application of the metaproteomic methods above (iv) in combination with metagenomics, in order to identify the full gene sequences coding for the induced novel cellulolytic enzymes (paper V)
3. Methodology

3.1 The construction of methanogenic microbial communities at strictly defined culturing conditions

Two different microbial communities were used in this study. A mesophilic microbial community was started from a full-scale biogas reactor that treated slaughter house waste at 38 °C, and a thermophilic microbial community was collected from a co-digestion plant that was operated at 52 °C.

Figure 5. Constructed methanogenic community. The constructed methanogenic community was started from a full-scale biogas digester and operated under fully defined culturing conditions in the laboratory. This way, it was viable in a clean environment that enabled the subsequent induction and identification of hydrolytic enzymes performed in later experiments.
Digestate from each reactor was transferred to a separate experimental biogas reactor [120] and kept under anaerobic conditions on chemically defined medium that provided all required nutritional ingredients and conditions at the respective temperature. The chemical medium that was used in this study is a result of the combination of several other recipes [121-123]. It contains over 40 individual components and was refined several times during this study to allow for the microbial communities to reach a metabolic steady state. Detailed information on the final recipe is given in paper I [118]. The gradual exchange of culturing medium during the experiments resulted in a microbial community in a constructed environment that was eventually free of components and impurities from the original biogas reactor habitats after an operation time of approximately five months.

3.2 Induction of targeted enzyme activity in the methanogenic microbial communities at a metabolic steady state

Over time, the exchange of the original environment for chemically defined medium in the experimental biogas reactor created an environment that was free of impurities from the original habitat as well as hydrolytic enzymes.

The term “metabolic steady state” is used in this study to describe an unchanging metabolic state that the microbial community reaches when it has been fed only the essential nutrients in their simplest, chemical monomeric form, e.g. glucose instead of polysaccharides, for many generations. In this state, all external factors are kept constant and the gas production from metabolizing the nutrients of the chemically defined medium is steady and reproduced in each feeding event. The secretion of hydrolytic enzymes in this state was found to be strongly suppressed. This suppression was expected, because if chemically simple nutrients are available, the energy and nutrient consuming task of producing hydrolytic enzymes will not be necessary and therefore not undertaken by the microorganisms of the microbial community by using evolutionary principles of microbial metabolism [124]. From this enzyme suppressed steady state, it was possible to control and induce targeted hydrolytic activity by adding the complex substrate during a starvation period [124] and to identify the microbial enzymes against a clear background (see paper I) [118]. For the induction of cellulolytic enzyme activity, cellulose in form of filter paper (Whatman no.1) was chosen as the sole complex carbon source instead of the simple monomeric sugar glucose.
For the induction of proteases, the protein bovine serum albumin (BSA) was used instead of free amino acids.

### 3.3 Enzyme activity assays

Enzymes catalyze the reaction from a starting substance called substrate to a product, and increase the rate of reaction by lowering the energy of the transition state. For the estimation of how much enzyme is included in a sample and to compare different enzyme samples and reaction conditions, it is common to measure the enzyme activity.

The activity of an enzyme is defined as the conversion of one mole substrate to product per second under strictly defined reaction conditions and the SI-unit used is katal (kat). Although, it is more common to use the old enzyme unit U, which is related to katal by the following: one international unit IU catalyzes 1 µmol substrate within 1 min of time = 1/60 µmol/s = 16.67 nmol and 16.67 nkat catalyze a rate of 16.67 nmol/s and therefore 1 U corresponds to 16.67 nkat [125]. If the substrate and the product differ in detectable features, the formation of this specific reaction product can be measured over time [126], according to the Nomenclature Committee of the International Union of Biochemistry (NC-IUM).

The enzyme assays used in the work described in this thesis detect the release of the fluorescent dye resorufin by hydrolysis of fluorogenic substrates. This release can be monitored by absorbance or fluorescence spectroscopy. The dye resorufin is released upon hydrolytic cleavage of the specific substrate used, caused either by protease or cellulase activity and the product exhibits an absorbance maximum at 573 nm and fluorescence emission 585 nm at pH 5.8. The release of the dye resorufin is monitored over time and by comparing the detected signal to a standard curve known concentrations, it is possible to calculate how many µmol of the substrate were cleaved per minute and per ml sample, which corresponds to the enzyme activity U.

#### 3.3.1 Protease assay

Proteases are hydrolytic enzymes that cleave peptide bonds in proteins. For detection of the proteolytic activity, the Universal Protease Substrate distributed by Roche was used, which consists of the resorufin-labeled protein casein as the substrate (figure 6). Upon hydrolytic activity, the protein is hydrolyzed into smaller peptides that still contain the dye resorufin. In a protein precipitation step,
the undigested protein is removed, the resorufin-labeled peptides stay in solution and can be detected by absorbance or fluorescence measurements.

Figure 6. Universal Protease Substrate. The protein casein is labeled with the fluorophor resorufin that is released upon proteolytic digestion and can be measured fluorometrically.

3.3.2 Cellulase assay

Upon cellulolytic activity, the polysaccharide cellulase is cleaved into smaller units by hydrolytic enzymes called cellulases. For the detection of cellulolytic activity, the Fluorescent Cellulase Assay Kit distributed by the company Marker Gene was used. The substrate for detection is the resorufin-labeled disaccharide cellobiose (figure 7) that releases the fluorescent dye resorufin upon cleavage of the β-1,4-glycosidic bond. Cellobiose is a substrate for cellobiohydrolases, but also exo- and endoglucanases have reportedly been detected [127]. In order to calculate the cellulolytic activity, the release of resorufin is measured constantly during a defined time period. By comparison to a standard curve of known dye concentration, the release of dye in µmol per minute and thus the enzyme activity can be calculated.

Figure 7. Resorufin-β-D-cellobioside. The synthetic substrate resorufin-cellobioside is used in the cellulase assay as fluorometric substrate. Upon cellulolytic activity, the fluorophor resorufin is released and can be spectrophotometrically measured and quantified.
3.4 2-D gel electrophoresis

Two-dimensional (2-D) gel electrophoresis is a separation technique that is used to separate complex protein mixtures. The basic principle is electrophoresis, which describes the migration of charged species in an electric field.

The method was developed by O’Farrell in 1975 [128] and consists of a protein separation step by isoelectric focusing (IEF) in the first dimension, followed by a second separation step by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), known as the second dimension. Thus, proteins are separated due to their isoelectric point and their molecular weight. The result is a protein SDS PAGE-gel, in which, ideally, each protein from the sample is represented by one distinct spot with a specific isoelectric point (pI) and molecular mass. The process is illustrated in figures 8 and 9. The protein distribution pattern can then be analyzed by special image analysis software. By comparing different 2-D gels, e.g. a treated and an un-treated sample, it is possible to detect changes in the protein expression pattern in the samples, e.g. up- or downregulation of proteins under certain experimental conditions. Interesting protein spots can, subsequently, be excised from the gel and further analyzed by e.g. Edman-sequencing or mass spectrometry to identify the proteins by their amino acid sequence.

In this project, 2-D gel electrophoresis has been used to analyze the extracellular metaproteome for the identification of proteins that were upregulated and secreted during the enzyme induction experiments. The key objective was to discover and identify enzymes that were produced, active and stable under the conditions in the microbial community of the biogas reactor.

3.4.1 Isoelectric focusing - the first dimension

Proteins are composed of different amino acids that are connected chemically by peptide bonds. The amino acids have different physicochemical properties and can have a positive or negative electric charge or be uncharged, depending on the pH of the solution they are dissolved in. For each protein, there is a distinct pH at which the sum or net charge of the protein is zero. This pH is known as the isoelectric point (pI) of the protein. When an electric field is applied to a gel matrix, that consist of a pH gradient and contains a protein sample, the protein will migrate to the point where its net charge is zero, which is the principle of isoelectric focusing (IEF).
When the IEF method was first introduced by O’ Farell in 1975, it was common to use IEF tube gels that were placed in a column. This approach had some drawbacks, as the separation was not stable over time and a drift of the proteins toward the cathode was observed [129]. The method was further refined by the introduction of immobilized pH gradient strips (IPG) for IEF [130,131], which led to more stable and robust IEF results and is also known as the standard method generally used by most laboratories today. Here, the IEF gel is immobilized on a thin plastic backing and is rehydrated prior to analysis.

An important step in the sample preparation is the complete denaturation and solubilization of the proteins to be analyzed. In order to ensure an efficient separation by isoelectric focusing, the proteins in the sample need to be denatured, reduced and fully solubilized. Only when the proteins are in an unfolded, disaggregated state, can the amino acid side chains interact with the gel matrix in the electric field and the separation due to the pI takes place. Another important aspect in sample preparation for 2-DE is the complete solubilization of the protein sample. Protein samples for 2-D gel electrophoresis need to be of high purity and contaminants such as lipids, nucleic acids, humic acids, salts or other charged species tend to cause severe interferences and incomplete separations [131]. This demand for a high purity requires effective purification methods that often involve protein extraction and precipitation steps. Moreover, pellets of precipitated proteins tend to be difficult to re-dissolve, thus efficient resolubilization buffers are needed.

To meet those requirements, different 2-DE sample buffers have been developed and they generally consist of chaotropic reagents such as high concentrations of urea or thiourea, zwitterionic or non-ionic detergents such as CHAPS or Triton X-100, reducing agents such as dithiotreitol (DTT) or tris(2-carboxyethyl)phosphine (TCEP) [132] and carrier ampholytes. The chaotropic reagents cause a disruption of intra-and intermolecular hydrogen bonds and thus denaturation and unfolding of the proteins. Hydrophobic interactions of different protein domains and thus aggregation and possible precipitation of proteins are prevented by the detergents. It is important to note that only non-ionic or zwitterionic detergents are used to avoid the introduction of extra charges to the denatured proteins. Reducing agents disrupt intra-and intermolecular disulfide bonds and are necessary for complete unfolding of the proteins and lastly, carrier ampholytes improve the overall solubility of proteins and aid in the focusing process [133].
For isoelectric focusing, the protein sample can be applied to the IPG strip in different ways. The most common used methods are passive/active sample rehydration loading, as well as paper bridge and cup loading. For the active or passive in-gel rehydration, the protein sample is added to the rehydration buffer that is used to hydrate the dry IPG strip before starting the experiment. Active rehydration involves the application of a low voltage to the gel in order to improve the entry of the sample into the IPG gel matrix. Passive rehydration describes the rehydration of the strip in presence of the sample without the application of a low voltage. In both techniques, the sample is soaked into the dry gel matrix.

For paper bridge loading, a thick piece of filter paper is moistened with the sample and placed between the rehydrated IPG strip and an electrode. Up to 500 µl sample can be applied to the IPG strip using this method [134].

For cup-loading, a small plastic container is carefully placed on top the rehydrated IPG strip close to an electrode. The cup usually has a loading volume of approximately 150 µl. Similar to paper bridge loading, the proteins enter the IPG strip gel matrix as soon a voltage is applied to the strip.

![Figure 8. 2-DE first dimension run](image)

**Figure 8. 2-DE first dimension run.** The protein mixture is applied to the IEF gel strip by anodic cup loading. The denatured proteins separate on the immobilized pH gradient gel by their net charge (pI) in an electrical field.

To start the protein separation by isoelectric focusing, the IPG strip needs to be placed into an IEF cell. In this cell, the low pH end of the strip is connected to the anode and the high pH end is connected to the cathode. A voltage gradient is applied at constant current and temperature conditions, the parameters are specific for each protein mixture and need to be determined empirically. The gradient usually starts at a low voltage to enable protein entry into the gel and to remove salts that quickly migrate toward the electrodes. This step is followed by a ramping step where the voltage is slowly raised to a maximum value in order to reach a high electric field strength. For each experiment, the complete amount of loading voltage is stated over time as applied volt hours (Vh). The separation of the proteins is dependent of the sample loading method, the point of sample
application (close to the anode or cathode), the gradient chosen for the voltage change and the overall time of focusing. The process is illustrated in figure 8.

3.4.2 SDS PAGE - the second dimension

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE) is a method used to separate proteins by their molecular weight in an electric field in a gel matrix. For this purpose, the proteins are treated with the strong detergent SDS that on the one hand denatures the proteins and forces the unfolding of the protein’s polypeptide chain into a linear form and on the other hand binds to the proteins backbone through its hydrophobic tail in a constant and sequence independent ratio of approximately one SDS molecule to two amino acids [135]. Since the sulfate group of SDS is negatively charged, the proteins are denatured and get a net negative charge. Additionally, in order to ensure a complete linearization of the proteins, reducing agents such as DTT or β-mercaptoethanol are added for breakage of disulfide bridges. The gel’s polyacrylamide matrix resembles a mesh and when the denatured, negatively charged proteins are applied to the gel and an electric applied through the gel, the proteins will migrate towards the anode. The protein sample is separated by the molecular weight while the proteins are pulled through the netlike gel matrix. Smaller proteins will migrate faster and larger molecules will stay closer to the application point, and as a result, the proteins will be separated by their molecular size.

In 2-D gel electrophoresis, the protein sample to be separated consists of the IPG strip after IEF. Prior to SDS PAGE, the strip is treated with a so called equilibration buffer that contains SDS, reducing agents and buffers. Subsequently and to enhance the separation, the strip is incubated with a second equilibration buffer containing iodoacetamide that alkylates free cysteines from disulfide bond reduction and thus, prevents a reoxidation and reformation of disulfide bridges during electrophoresis. The strip is then placed onto the gel using molten agarose. The agarose seals potential gaps, ensuring an even contact between the IPG strip and the SDS gel and a good migration of the proteins. The electrophoresis is usually started at a low voltage to allow an even entry of the proteins from the IPG strip into the SDS gel. The voltage is then increased, depending on the gel length, thickness and composition. The separation time is also depending on those parameters. To follow the electrophoresis, a tracking dye is added to the sample and the addition of marker proteins (protein ladder) allows to estimate the molecular weight of the separated proteins (figure 9).
Figure 9. 2-DE second dimension run. The IEF strip containing separated proteins by their pI is applied to a SDS-PAGE gel after a treatment with iodoacetamide and SDS containing buffer. By applying voltage, the negatively charged proteins separate by their molecular weight in the gel matrix.

After completed electrophoresis, the IPG strip is removed and the gel is stained to visualize the separated proteins. To prevent diffusion of the sharp protein spots over time, a fixation step is performed using acetic acid and ethanol. The visualization can be achieved by staining the gel with different dyes, e.g. Coomassie Brilliant Blue dyes, silver stains or fluorescent dyes, such as SYPRO Ruby. The stains have different advantages and disadvantages and also differ in handling, sensitivity, necessary equipment for visualization and compatibility with downstream mass analysis.

3.4.3 Image analysis

After electrophoresis and staining, an image of the 2-DE gel is captured for documentation and digitalization. This can be done either by using a scanner or a CCD (charge-coupled device) camera system. The gel is further analyzed using an image analysis software for 2-D gel electrophoresis. The software runs a spot detection algorithm that discriminates the spots from the gel’s background and allows a quantification by spot size, intensity and density. Distortions in the gel and differences in the run are automatically corrected or can be edited manually. In 2-DE experiments, it is common to compare two similar samples originating from the same or very similar environment, so the gels will therefore share many proteins. Using an image analysis software, it is possible to perform a gel-to-gel comparison and spot matching can be made in order to identify proteins that are shared, up- or down-regulated (seen by differences in density/intensity) or proteins that appear or disappear. Different statistical analysis methods can be performed, once all spots are recognized and matched.
After protein identification, the interesting protein spots are usually excised from the gel, treated with a site-specific protease and identified by mass spectrometry [136].

3.4.4 2-D DIGE

In 1997, Ünlü et al. introduced the method of two-dimensional difference gel electrophoresis (2-D DIGE) [137], which circumvents problems with gel-to-gel differences in 2-DE. 2-D DIGE allows for the separation of multiple samples on the same 2-DE gel by labeling the samples with different fluorescent dyes. This makes it possible to avoid common problems that occur when comparing different gels, e.g. differences in running conditions during 2-DE or slight differences in the composition of the gel matrices. The proteins are covalently labeled with distinct cyanine dyes that emit light by fluorescence at different wavelength prior to IEF. The dyes have similar physicochemical properties and do not introduce significant differences in the pI or molecular weight of the labeled proteins. In this way, the differently labeled samples will co-migrate during the 2-DE run, identical proteins will end up in the same spot on the 2-D gel and image analysis (spot matching) will be improved and more reliable.

In the most popular protocol, three different fluorescent Cy-dyes are used for DIGE: Cy2, Cy3 and Cy5. These are available as N-Hydroxysuccinimide (NHS) esters, which form an amide bond with the ε-amine group of lysine in the protein. The labeling reaction is performed as a so called minimal labeling procedure, where approximately only 3% of all lysines in the protein become labeled. This strategy still allows for a high detection sensitivity comparable to silver staining [138] and avoids chemical modifications on all lysine residues that would change the properties and structure of the proteins. The technique is therefore compatible with down-stream analysis, e.g. gel electrophoresis and mass spectrometry analysis. After gel electrophoresis, an image for each dye is captured by either using a high quality laser scanner or CCD camera that is equipped with appropriate fluorescence filters. An image analysis software for DIGE is then used to merge and analyze the different channels, thus information about the protein distribution and changes in intensity (up-or downregulation) can be generated. It is furthermore possible, to display the different fluorophores with a distinct color, so that differences in the protein expression can be easily spotted in a false color picture (figure 10).
Figure 10. **2-D DIGE of a cellulase induction experiment.** The colored arrows represent the different conditions and samples in the induction experiment. The red arrow stands for a cellulase induced and -active sample (4 d), blue for the un-induced control at the same time and green for an early/inactive time of induced sample (1 d). The colors match the spots on the 2-D DIGE gel. Red spots correspond to extracellular proteins that are upregulated due to the cellulase inducing event, blue spots are from the uninduced/inactive control sample and green spots from the inactive induced early sample. If spots are present in several samples, a mixed color is shown.
One advantage of the DIGE methodology is the separation of different samples on the same gel and thus, improved image analysis due to perfect spot matching and elimination of gel-to-gel variability. Another advantage is the use of the Cy2-channel as the so called internal standard [139]. Using this internal standard technique, all samples that are investigated in an experiment are also labeled with Cy2 and mixed in equal partial amounts, resulting in an internal standard. Additionally to the Cy3 and Cy5 labeled samples, the internal standard is run on every gel in the experiment. Since the internal standard always contains the same amount and species of proteins, all gels in the experiment can be easily matched, compared and normalized to each other, even if they were not separated on the same gel. This technique increases the reliability of spot matching and also reduces the number of gels needed to detect statistically significant differences between samples.

For the experiments discussed in this thesis, the secreted proteins of a cellulose-induced microbial community were compared to the secreted proteins of a non-induced sample by 2-D DIGE. In this way, the dynamic changes in the protein expression pattern could be accurately monitored and linked to the inducing substance, thereby allowing the detection of enzymes that are upregulated, secreted and involved in the degradation of cellulose, see figure 10. Details are presented in paper IV and paper V.

3.5 Zymography

Zymography is an analytical detection method that allows for the visualization of hydrolytic active enzymes by conversion of their specific substrates. The visualization of activity can be accomplished either by the disappearance of the substrate or the appearance of a degradation product. Zymography can be used in electrophoresis gels (in-gel zymography), tissue samples (in situ zymography) and in intact organs (in vivo zymography) [140].

In the work presented in this thesis, the presence of cellulases and β-glucosidases in the extracellular environment of a microbial community was visualized by in-gel 2-D zymography.

For in-gel zymography, the sample is separated by SDS-PAGE or native PAGE under non-reducing or reducing conditions, depending on the enzyme and the specific method used. The enzyme’s substrate can either be incorporated in the gel-matrix, added in a staining buffer or the gel can be blotted on membranes and
overlaid with agarose containing the substrate [140]. After electrophoresis, the SDS and possible reducing agents need to be removed to allow the enzymes to refold to a hydrolytic active form, which is usually achieved by extensive washing steps using an appropriate buffer. The washing is followed by incubation in an activity buffer that establishes the conditions the enzyme needs for activity, e.g. by establishing the optimal pH. Depending on the approach used, the non-converted substrate can either be stained using an appropriate dye or the degradation product can be visualized by e.g. fluorometric methods. The results are then documented with cameras, scanners or light tables that are commonly used for electrophoresis experiments.

For 2-D in-gel zymography the proteins in the sample are separated by IEF and SDS PAGE prior to activity staining. It is important to note that the alkylation step of cysteins, which is generally used to achieve an improved resolution of the second dimension, is omitted, since it hinders refolding of the enzymes to an active state.

3.5.1 In-gel zymography of cellulases using Congo Red activity staining

In order to detect possible cellulolytic activity, the extracellular samples were analyzed by 2-D in-gel zymography using the water soluble endoglucanase substrate carboxy-methyl cellulose (CMC) incorporated in the 2-D gel matrix [141]. The diazo dye Congo Red has a strong binding affinity to cellulose, and will bind to the non-degraded cellulose matrix upon incubation on the zymogram gel. Thus, protein spots that show cellulolytic activity contain no CMC and appear as colorless halos against a deeply red stained gel background.

3.5.2 Visualization of β-glucosidases by 4-MUC staining

The presence of β-glucosidases can be detected by use of the substrate 4-Methylumbelliferyl β-D-cellobioside (4-MUC) [142]. Upon glucosidase activity, the dye 4-methylumbelliferone is released, with an absorbance maximum at 350 nm and a fluorescence signal that can be detected at 455 nm. When used for detection in electrophoresis gels, the dye is added to the activity buffer and directly incubated with the gel. Upon illumination with UV-light and use of a fluorescence filter at 455 nm, β-glucosidases appear as illuminating, fluorescent bands or spots on the gel.
3.6 Tandem mass spectrometry

Mass spectrometry (MS) is an analytical method used to determine the mass-to-charge (m/z) ratio of ions in a vacuum, thus, enabling the determination of the molecular masses of e.g. an investigated protein sample. In its simplest form, a mass spectrometer consists of three major parts: the ionization source, the mass analyzer and the detector. Upon analysis, the sample is ionized typically by addition of protons, subsequently separated by the ions mass-to-charge ratio in the mass analyzer and finally registered in the detector [143]. If two mass spectrometers are combined, it is possible to perform fragmentation experiments that allow to partly resolve the amino acid sequence of the protein [144]. This method is called tandem mass spectrometry (MS/MS). The information obtained can be used to identify the proteins by matching the detected or theoretical peptides to known proteins that are available in public protein databases [145]. It is further possible to identify potential post-translational modifications (PTMs) and also to quantify proteins in specially designed experiments [146, 147].

Starting in the late 1980s, the development of soft ionization methods such as electrospray ionization (ESI) [148] and matrix assisted laser desorption/ionization (MALDI) [149] enabled the analysis of the large biomolecules polypeptides and led to the development of highly sensitive and accurate multistage mass spectrometers [145]. The combination of two mass analyzers allows for a selection and fragmentation of specified peptides called precursor or parent ion, thus causing a breakage of the selected peptide usually along the peptide’s backbone and resulting in smaller parts (product ions). The fragmentation can be achieved using different techniques, e.g. collision induced dissociation (CID) as in the Orbitrap MS, the method used for the work presented in this thesis, which is performed in a so called collision cell using an inert collision gas. During this fragmentation process, the peptide backbone is broken ideally along the amide bonds and b- and y –product ions are created, named depending on the position of the remaining charge (see figure 11).
Figure 11. Nomenclature for MS/MS peptide fragmentation. Upon collision with an inert gas, the peptide backbone is fragmented into several pieces, so called product ions. Depending on the fragmentation technique, different charged product ions are formed. The subscript specifies the number of residues in the fragment. If the charge is retained on the N terminus of the peptide, the ion is classed as either a, b or c. If the charge remains on the C terminus, the ion type is labeled x, y or z [150].

In the recorded tandem mass spectrum, the m/z difference between the particular peaks allows for the determination of each amino acid position and thus partial conclusions on the sequence. Additionally, the overall fragmentation pattern is recorded. Thus, a typical MS/MS experiment delivers three kinds of information: (i) the molecular weight of the parent ion, (ii) the partial amino acid sequence of the further fragmented ion and (iii) the overall fragmentation pattern [151]. The combination of data allows for a more precise analysis and identification of proteins as compared to simpler MS experiments.

When performing a typical protein identification experiment using MS/MS, the protein sample is initially purified, the individual proteins are separated and subsequently digested with a highly site-specific protease (e.g. trypsin) which leads to structurally well-known peptides. In case of trypsin, amide bonds are cleaved after the amino acid arginine (R) or lysine (K). Since small fragments are used to identify the bigger protein, this approach is also called bottom-up proteomics [152].

After data acquisition using MS/MS, there are several different algorithms available for identifying the protein using the fragmentation data and protein databases. The approach used in this thesis is illustrated in figure 12.
Figure 12. Workflow of proteomic identification by 2-DE and MS/MS. Proteins are separated by 2-DE and analyzed by image analysis. Enzyme candidates are picked from the gel and subsequently trypsinated resulting in defined peptides. The peptides generated are fragmented and sequenced in a MS/MS experiment and finally used for database searching.

One approach of data analysis is the comparison of the fragmentation pattern, used e.g. in the software packages SEQUEST, MASCOT [153] and X!Tandem [154]. The recorded fragmentation pattern of the parent ion is compared to a theoretical in-silico fragmentation pattern of database protein entries that show the same molecular weight as the precursor ion after a theoretical digestion of the database entries with the same site-specific protease as used in the experiment. By this comparison the amino acid sequence of the analyzed peptide can be obtained.

Another approach is the combination of amino acid sequence data derived from the fragmentation spectrum and the mass data in so called “sequence queries” or “sequence tags” [155]. The described approaches are efficient if homologs to the analyzed proteins can be found in public protein databases, but in cases when there are no exactly homologous sequences available, as is often the case in metaproteomics studies of microbial communities, it is more advantageous to also
use the amino acid sequences that can be extracted from the fragmentation spectrum itself, by so called de novo sequencing.

The software packages Lutefisk [156] and PEAKS [157] for example make use of this de novo sequencing approach, where the amino acid sequence is calculated from the MS/MS spectrum and then compared to the database entries.

All algorithms and software packages subsequently use scoring and correlation values to evaluate and rank the matching to the entries found in the database. An important parameter is the calculation of the false discovery rate (FDR) that describes the accuracy of a measurement by giving a percentage value of false positive identifications. One way to calculate the FDR is the target-decoy approach [158] as used in many analyzing software packages. In this approach, a reversed or randomized database of the same size as the database used in the experiment is created and searched against the experimental data using exactly the same experimental settings. Found matches in this decoy data set are inevitably false, thus enabling the calculation of the FDR in the whole dataset and giving an estimation for the quality of the matching.

Another important value used to judge the quality of the identification, is the probability value p, which is defined as the probability that a false identification in a search achieves the same or a better matching score as an actual hit [159].

The different software for peptide identification further refine the p-value to internal scores, e.g. the software PEAKS, used in papers IV and V, converts the p-value to a ranking peptide score by calculating the -10*\log_{10}(p-value), so that the higher the score, the lower the p-value and the more reliable is the identification. This way, the hits can be ranked accordingly to their match quality. PEAKS also includes an algorithm called SPIDER that allows for the detection of peptide mutations and thus performs cross-species homology searches [160].

For the experiments performed in papers IV and V, tandem mass spectrometry by LTQ-Orbitrap was used to identify the proteins judged as upregulated from a 2-D DIGE gel upon induction of extracellular hydrolytic enzyme activity in microbial communities at a metabolic steady state. For protein identification, the traditional database approach, using public protein databases, and a metaproteogenomic approach by the use of a customized metagenome database of the microbial community from which the upregulated proteins originated, were used. The approaches are described in details below.
3.6.1 Database searching in tandem mass spectrometry for protein identification

The most common purpose of MS/MS experiments in metaproteomics is the identification of protein function or microbial origin of the investigated protein samples for understanding physiology or microbial ecology. This is usually achieved by comparing the detected peptides, peptide sequences or peptide sequence tags to proteins in databases.

The largest amounts of publicly available DNA and protein sequences is available from three major databases that share their data on a daily basis, specifically Genbank at the National Center for Biotechnology Information (NCBI) in Bethesda/USA, the DNA Database of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL) at the European Bioinformatics Institute (EBI) [161]. Additional to these major databases, there are other databases that vary in size and quality of the contained data.

For mass spectrometry, examples for frequently used universal protein databases are NCBI’s nr (non-redundant) database and UniProt, which combine several other databases.

The NCBI nr database for example is assembled by the NCBI as a protein database for BLAST searches, a tool that allows for the identification of proteins due to sequence similarities. The nr database is a collection that contains non-identical sequences from other databases including as GenBank, PDB, Swiss-Prot, PIR, and PRF [162, 163], which will not be reviewed in detail here. It is important to note, that by combining large amounts of data, the nr database is broad and comprehensive, but also very large [162]. During the writing of this thesis in November 2016, the nr database was a 55 GB large file that contained approximately 95 million entries. Another example is the UniProt database that contains high quality sequences from SwissProt, TrEMBL and PIR and is smaller and well curated [164]. In addition to these databases that contain overall sequence data, there are other databases accessible that contain only certain information about selected proteins, e.g. the CAZyme database mentioned earlier.

For the purpose of protein identification, it is important to note that by using a public database in an MS/MS experiment, the matching of the experimentally observed peptides can only be made to the sequence that is actually present in the database used. This is usually not a problem if the sample to be analyzed originates from a well-known and well-sequenced model organism, e.g. human, mouse or E.coli. It can then be expected that a high number of obtained peptides

44
from the protein sample can be matched to the corresponding database entry, making an identification very likely. However, if non-model organisms or organisms whose DNA has not been sequenced yet are the source of the protein to be investigated, it can be assumed that there is no sequence homolog present in the protein database. A low number of peptides may be matched to homologous sequences of evolutionary related organisms, especially if the peptide hits originate from highly conservative protein regions that are preserved among similar species. This imprecise sequence matching may not be a problem if the focus is on the identification of protein function, as in the majority of current metaproteomics studies. In these cases, it is sufficient that the MS/MS results consists of peptides matching to conserved regions that define the protein function. For the discovery of novel proteins for practical applications however, it is crucial that the actual protein sequence is obtained, so that the gene can be cloned for the production and characterization of the identified protein. This obstacle in protein identification can be circumvented by combing proteomic and genomic data from the same experiment, as in proteogenomics or metaproteogenomics approaches, which are explained in section 1.4.4.
Biotechnology has been defined by several governments and the EU commission as a key enabling technology for economic growth in a low-carbon and knowledge-based bioeconomy. The discovery of new enzymes is an important step to fulfill the environment friendly and sustainable development to a future industrial biotechnology. With this work, a significant contribution has been made to the scientific field of bioprospecting by development of a method that can be used to discover novel enzymes from microbial communities. The method is important, because microbial communities still represent a mainly unexplored source of novel enzymes. There are various reasons that metaproteomics has not yet been used for targeted enzyme discovery. One reason is the sheer magnitude of the task, which includes dealing with hundreds of thousands of different genes and the fluctuating expression of the corresponding proteins. At present, secreted, hydrolytic enzymes are not accessible from natural habitats using established metaproteomics methods, due to substances, which are present in the environment and interfere with sample preparation and downstream analysis. In addition, missing genetic information in public databases impedes the correct identification of proteins and the possibility to clone the correct gene.

Thus, to enable the use of metaproteomics for targeted enzyme discovery in microbial communities, the approach presented in this thesis is based on providing the necessary conditions for this purpose. Specifically, conditions were established that allowed for tight control of enzyme expression and sample preparation of the enzymes in the extracellular environment, together with the
different techniques that are presented in five papers: A methanogenic microbial community was kept viable at a metabolic steady state that allows for the induction of a targeted enzyme activity (paper I). The induced enzymes (cellulases) were shown to significantly improve the biomethane production rate and yield from the induced enzymes substrate (lignocellulose) experiment (paper II). A sample preparation method has been established that allows for the analysis of the specifically induced and upregulated extra-cellular enzymes by 1-D and 2-D gel electrophoresis experiments (Paper III). This enabled the selection of upregulated proteins against a large background of other proteins by differential 2-D gel electrophoresis and image analysis together with the identification of the function of these upregulated proteins by tandem mass spectrometry and database searches (papers IV). And finally, by using the same approach enhanced by the sequenced metagenome of the same microbial community that produced the enzymes, establishing a direct link between the selected induced enzymes of the metaproteome and the corresponding genes in the metagenome (paper V). This approach enabled the identification of both the function of the enzymes and their corresponding gene sequence that allows for molecular cloning of the identified genes. The papers are presented in detail in the following section.

**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 study was to develop a method that enables the controlled expression of targeted genes to allow for the identification of targeted extra-cellular enzymes from a microbial community by applying general principles of microbial metabolism.

The model microbial community was maintained on a chemically defined synthetic medium under strictly controlled culturing conditions. After a maintenance time of approximately 5 month, the microbial community in the constructed environment was eventually free of components and impurities originating from the full-scale biogas reactor inoculum. This clean and stable environment was established through the gradual exchange of culturing medium during the experiment and the simultaneously continuous consumption and/or wash out of complex organic material. By maintaining the microbial community viable and functional on a chemically defined medium with easily accessible
nutrients, the secretion of extra-cellular enzymes was found, as expected, to be strongly suppressed. This reduced activity thus provided a ground state or baseline expression of background enzyme activity for the induction of targeted extra-cellular enzymes. The induction was achieved, by removal of one highly accessible nutrient and simultaneous replacement by its complex counterpart, e.g. cellulose instead of glucose. This way, the members of the microbial community were “forced” to secrete hydrolytic enzymes in order to break down the complex molecules to make the nutrient available in its simple form. The thereby induced enzymatic activity could be detected in comparison to the low background enzyme activity in the pre-induced state and the method provided distinguishable differences in the protein expression pattern, before and after induction. Using these principles of microbial metabolism, it was possible to induce and detect cellulase and protease activity. Since most enzymes used in industry to date were discovered from pure-cultured organisms under laboratory conditions, the use of a microbial community for enzyme discovery experiments offers the possibility to detect new, unknown variants of hydrolytic enzymes that stem from the majority of microorganisms that are not available when using the established pure culturing techniques. Furthermore, since these enzymes are active and stable in the anoxic environment of a biogas producing community, the possibility to identify novel enzymes from the environment to which they are evolutionary adapted, is a valuable and promising approach. This is in contrast to commercially available enzymes that are derived from microorganisms that are not present in the anaerobic environment and not very suitable for enhancing biogas production.

**Paper II: Enhanced biomethane production rate and yield from lignocellulosic substrate by *in situ* anaerobic digestion treatment with endogenous cellulolytic enzymes**

In paper II, the effect of an induced cellulolytic enzyme mixture, obtained from the metabolic steady state reactor described in paper I, was investigated in a biochemical methane potential (BMP) test using lignocellulosic material as a substrate. The aim of the study was to investigate if the filter paper induced extracellular cellulolytic enzymes had any effect on their natural lignocellulosic substrate in a biogas experiment under standardized conditions. The addition of hydrolytic enzymes to increase the rate and yield in biogas production is still a rather unexplored field and cellulolytic enzymes cocktails that are commercially
available today are produced predominantly from aerophilic fungi that are not naturally present in this anaerobic environment. It can be assumed that those enzymes are not evolutionarily adapted to be active and stable under the prevailing conditions in a biogas digester and it is unlikely that foreign enzymes display the efficiency and life-time needed to be effective and profitable in the biogas production process.

Lignocellulosic substrate routinely used in biogas production has a low biological availability and the biogas production process could benefit from employing efficient hydrolytic enzymes to speed up the hydrolysis step. In biogas production, the hydrolysis step, in which complex, recalcitrant nutrients are broken down to their simpler chemical form, is generally seen as the rate limiting step and the addition of hydrolytic enzymes that perform well under the given conditions is an important target. The discovery of effective hydrolytic enzymes that enhance the hydrolysis step is therefore valuable and it is reasonable to search for those enzymes in the biogas environment itself.

In paper II, it could be shown that the addition of endogenous cellulases led to a higher biogas production rate and yield in the BMP experiment. The added cellulolytic enzyme mixture also showed an effect over a long period of time, essentially until the cellulosic substrate was entirely consumed. The result indicates that the induced enzymes in the extracellular fluid are persistent to the prevalent environmental conditions, thus forming a solid base for the continued analysis and identification of these induced enzymes.

**Paper III: Assessment of sample preparation methods for metaproteomics of extracellular proteins**

Paper III presents a systematic evaluation of different protein extraction methods for the analysis of the exometaproteome of the microbial community maintained in a constructed environment for 1-D and 2-D gel electrophoresis. The effects of sample concentration by lyophilization, of eight different protein extraction and precipitation protocols, as well as the influence of different sample solubilization buffers were investigated. The findings are important to allow for the reliable identification of the induced and secreted hydrolytic proteins from the methanogenic microbial community described in paper I and II.

The analysis of secreted proteins from microbial communities is hampered by a low protein concentration and presence of background contaminations, e.g. intracellular proteins, proteins of the surrounding environment or contaminants.
like humic acids that are known to interfere with subsequent analysis methods. Many proteomic and most metaproteomic methods conducted to date focus on the investigation of intracellular proteins in microbial ecology and physiology. There are, thus, no studies available that explore the secreted proteins of natural samples of complex microbial communities. As a result, the findings in paper III reveal that the application of commonly used sample preparation methods in proteomics, such as sample concentration by lyophilization and protein precipitation using the typical trichloroacetic acid (TCA) precipitation method, are inferior as compared to combinations of extraction and precipitation methods. The extraction of the exoproteome sample with a phenol-based buffer followed by a protein precipitation by using a methanolic ammonium acetate solution gave the best results in terms of spot quantity and separation in electrophoresis. The use of thiourea containing 2-DE sample solubilization buffer also resulted in an improved performance as compared to solely urea based sample buffer.

**Paper IV: Metaproteomics-guided selection of targeted enzymes for bioprospecting of mixed microbial communities**

The identification of cellulases from biogas producing microbial communities is an interesting target for enzyme discovery and biotechnological applications. Enzymes that are present in distinct environments are evolutionary adapted to be stable and active under the prevailing conditions. Lignocellulosic biomass is a natural, but highly recalcitrant renewable resource used for biogas production and a more efficient hydrolysis of the biomass by addition of stable cellulolytic enzymes would contribute to reach higher yields in the biogas production process.

Paper IV describes a method that is capable to induce a cellulolytic activity in a mesophilic biogas-producing microbial community at a metabolic steady state and to identify the proteins that are upregulated and secreted by 2-D DIGE followed by tandem mass spectrometry. By applying the knowledge presented in papers I, II and III, we were able to identify several carbohydrate-active enzymes. Of 39 investigated proteins 28 (72%) were found to be non-cytosomal and 17 (44%) could be annotated as carbohydrate-active enzymes. Thus, demonstrating both the applicability for identifying extracellular proteins and for guiding the selection of proteins towards those that were specifically upregulated and targeted by the enzyme inducing substance. In addition, since the identification of interesting proteins was based on the regulation of enzyme expression in response
to a need to hydrolyze and utilize a specific substance, also other unexpected enzyme activities were upregulated and identified. A database search using the SPIDER tool for the discovery of peptide mutations and cross-species homology searches revealed that the detected protein database hits were inaccurate and the accurate amino acid sequence of investigated proteins remains unknown. Thus, for the purpose of bioprospecting for enzymes to clone, produce and characterize, it was concluded that an identification against public protein databases was not sufficient to identify the correct gene or protein sequence for cloning of identified enzymes.

**Paper V: Metaproteogenomics-guided identification of novel cellulases by substrate induced dynamic gene expression in a microbial community**

In paper V, the induction process on a thermophilic methanogenic microbial community was investigated using a proteogenomic method for the identification of the precise gene sequence of secreted, stable cellulases at a temperature of 55 °C for enzyme discovery. Thermostable enzymes are of special value for industrial applications, since they naturally possess a high resistance to the surrounding environment even at moderate conditions, e.g. higher tolerance towards temperature, organic solvents, extreme pH ranges etc. Therefore, the discovery of new, unknown, thermostable enzymes is an important and crucial target for the development of efficient and environmental friendly bioindustries.

Using the same approach as described in paper IV, the secretion of cellulolytic enzymes could be induced by applying general principles of the microbial metabolism and subsequently identified by a combination of activity measurements, 2-D DIGE and mass spectrometry methods. As a complement, a customized protein sequence database was generated using next-generation DNA sequencing of the microbial community and thus applying a proteogenomics approach by combining techniques known from the fields of proteomics and metagenomics. Using this proteogenomics technique, it was possible to identify the precise gene sequence of the actual peptides and proteins present in the sample, not merely sequence homologs to proteins known from public DNA databases, thus providing the possibility for enzyme production. This is important, because the DNA and protein information available at public databases to date, do not
contain enough information from microbial community samples, thus hindering correct protein identification.

This approach differs tremendously from solely metagenomics or metatranscriptomics experiments that focus on the analysis of all genes or mRNA present at a certain point in time in a microbial community and need labor-intense analysis of all expressed genes or mRNA. In the approach presented in paper V, the community reacts to the altered nutrient conditions and responds by secreting cellulolytic enzymes. This dynamic response is visualized and analyzed on 2-D DIGE-gels and the proteins selected for MS/MS-analysis from this event should be closely related to the induction event only. By identifying the peptides, and thus the proteins, from the carefully chosen, upregulated spots, it is possible to narrow down the analysis process and to directly get an answer as to which cellulolytic enzymes are present and active at the chosen experimental conditions.

Using the presented approach, we were able to identify 23 spots of a total of investigated 24 spots as carbohydrate-active enzymes (hit rate 95.8%) and further to determine the complete amino acid sequence of four novel enzymes that showed homology to different glycoside hydrolases (hit rate 16.7%).

The results presented in paper V thus contribute to the efficient discovery of novel cellulolytic enzymes from a biogas producing microbial community at thermophilic conditions.
5. Conclusions

With the work presented in this thesis, I have successfully contributed to the development of a metaproteomics-guided method for the highly targeted induction and correct identification of extracellular hydrolytic enzymes from microbial communities. Since the exact gene sequence of the newly found enzymes is given using this approach, a production of exactly these enzymes with the desired properties is possible by molecular cloning methods. Thus, the method allows for effective bioprospecting for novel enzymes of industrial interest in microbial communities.

The developed method is important, because microbial communities present a highly unexplored field for bioprospecting of novel enzymes, due to their natural high diversity and the abundance of unknown enzymes. Recent progress in the bioanalytical fields of –omics techniques and the emergence of microbial community focused meta-omics techniques have led to an increased research activity in terms of bioprospecting, but suffer from a very low hit rate and the lack of genetic information for sequence homology-based identifications. The newest developments by combining different –omics approaches, so called integrated –omics, are very promising in overcoming these limitations. However, to date they have only been used in microbial ecology and medical studies as well as bioremediation projects, and not for the purpose of enzyme discovery for industrial biotechnology.

The metaproteomics-based method for the rational induction and identification of interesting enzymes from microbial communities presented in this thesis, bypasses the limitations of the individual meta-omics approaches. By developing a
culturing technique that allowed for the maintenance of the microbial community at a metabolic steady state, it was possible produce and identify enzymes under the conditions and substrates of choice. Furthermore, an environment was created that was free of impurities from the original habitat as well as from hydrolytic enzymes. This allowed for the identification of upregulated extracellular proteins by the enzyme induction experiment against a low background of extracellular proteins without further complication by analysis-interfering substances, e.g. humic acids and fats. This is a key aspect for the successful investigation of the induced dynamic change in the protein secretion pattern, since other metaproteomic investigations suffer e.g. from the presence of down-stream analysis interfering substances or a low resolution of investigated electrophoresis gels. These targeted, upregulated extracellular proteins could be visualized and measured by application of the sensitive, advanced gel electrophoresis-based proteomic technique 2-D DIGE and led to a precise selection of protein candidates for further investigation by tandem mass spectrometry using a metaproteogenomics approach. This approach uses the six-frame translated metagenome of the microbial community as a database for peptide matching and led to the precise identification of the gene sequences of the proteins investigated. This way, the method enables the identification of the exact gene-sequence of proteins that are expressed due to the altered culturing conditions. It can be concluded that the enzymes found are most probably involved in the breakdown of the new substrate and are therefore interesting candidates for further characterizations. Since those proteins originate from a methanogenic microbial community, they could be expected to be stable in the environment of other methanogen communities e.g. in biogas digesters, especially in comparison to “foreign” fungal enzymes that are industrially produced and used today.

Since the approach is highly targeted, the number of proteins for characterization in rather low as compared to other approaches and is typically between 4 to 8 candidates that need to be further investigated out of 30 to 40 proteins analyzed. In the initial experiments for the identification of novel cellulases, the hit rate for the functional identification of the targeted, upregulated proteins was between 44% and 96%. In the final metaproteogenomic experiment, it was possible to identify the exact gene sequence of four different novel cellulases. This genomic identification resulted in a hit rate of 16.7%, which is notably higher than in established meta-omics approaches that show a hit rate of 0.083% [54]. Our results indicate that the hit rate could be even higher, if the
quality of the metagenomics database increases. Due to current technical limitations in next-generation DNA sequencing (NGS), the metagenomic database contains sequenced DNA fragments (contigs) of differing length. In order to truly identify and to produce a gene, the identified peptides need be matched to a complete gene sequence. In our first metaproteogenomics experiment, this was the case for 4 investigated protein spots on the 2-D DIGE gel. The remaining spots contained peptides that were matched to incomplete genomic sequences. However, this problem is caused by the current technical limitation of NGS sequencing that results in a high number of short DNA fragments and is not due to the feasibility of the approach. Future progress in NGS techniques will result in longer DNA reads and assembled sequences (contigs), and thus it can be concluded that an increased number of longer DNA fragments enables an improved identification of complete genes in metaproteogenomic studies.

An important difference to other meta-omics approaches is, that the selection of possible enzyme candidates is done by identifying upregulated proteins on the 2-D electrophoresis gel and not due to a predicted sequence homology. Therefore, the selection may also lead to the identification of enzymes that were earlier not known to participate in the degradation of the substrate. This is a very interesting result that not only helps to understand the complex degradation processes in microbial communities but also enables the identification of completely unknown proteins that were not known to participate in the degradation. This possibility for increased identifications is very beneficial for efficient bioprospecting for novel enzymes.

Another very interesting characteristic of the technique presented in this thesis is the possibility to recognize enzymes that may work complementary and synergistically in the hydrolytic decomposition of the substrate used in the induction experiment. By studying the protein expression pattern on the 2-D DIGE gel, it is possible to recognize the proteins that are expressed at the same time by the microbial community in response a certain stimulus. For the efficient degradation of lignocellulosic material for example, it is known that the presence of different carbohydrate-active enzymes of varying enzymatic activity and substrate-specificity is needed. An accurate analysis of the 2-D DIGE gel in combination with the metaproteogenomics approach gives information about the identity and quantity of the expressed enzymes, which allows for the composition and production of so called enzyme cocktails that contain mixtures of different
enzymes and are increasingly used in the industrial decomposition of lignocellulosic biomass.

It is important to emphasize that the focus of this work was not to contribute to the understanding of the microbial composition or metabolic pathways of the investigated microbial community, but to enable the precise induction and effective identification of the gene sequence of extracellular hydrolytic proteins that were expressed under certain stimuli. The value of this method is that different microbial communities could be investigated and used for enzyme production under substrates and conditions of choice. Using the developed technique, our group has been able to induce different hydrolytic enzymes, specifically cellulases, chitinases and proteases in a mesophilic and thermophilic methanogenic microbial community. In future projects, it would be interesting, to induce other interesting enzyme activities, e.g. amylases or lipases that are of growing interest for industrial biotechnology. But also the investigation of other microbial communities, e.g. originating from extreme habitats, would be interesting, since they possess especially physicochemical stable proteins. A different future project would be to change the external stimulus in order to induce enzyme production. If the stimulus is changed from a nutrient substrate to e.g. a toxin, the response of the microbial community to the hazardous conditions could be studied and result in novel enzymes that may be used in bioremediation projects for the effective removal or neutralization of pollutants in contaminated places.
Acknowledgements

My time at IFM comes to an end and writing these acknowledgements is the finishing touch on my thesis. It has been a period of intense learning, not only in the scientific arena, but also on a personal level. Moving to a foreign country, starting as a PhD student, meeting many wonderful, talented persons and writing this thesis had a big impact on me.

I would first like to thank my supervisor Martin Karlsson for the enthusiasm, support and continuous optimism about this project. I consider it a privilege to have been your first graduate student. I appreciate your brilliant ideas and immense knowledge that helped me to grow as a research scientist.

I would like to extend my thanks to my co-supervisor Nalle Jonsson, who first accepted me as a diploma student and later as a PhD student. Thank you for always having an open door and advising me throughout the years. You have the remarkable talent to remember even the tiniest details of my experiments and still pay attention to the big picture. In addition, I would like to thank Uno Carlsson for valuable advice during the initial period of this project and much appreciated comments on our publications. My deepest gratitude to Hans-Lothar Fuchsbauser for introducing me to the fascinating world of enzyme technology and for encouraging me to start my PhD studies. I would like to acknowledge InZymes Biotech AB for financial support and Tekniska verken i Linköping AB for technical assistance in running and maintaining the biogas digesters.

To Maria Turkina, thanks you for the support with the MS experiments and Bijar Ghafoori for help and advice on the initial 2-DE experiments.
I have been very fortunate to work in a group that holds many wonderful persons: My fellow PhD student Mikaela Johansson. Things improved so much when you joined our group! Thank you for the support, nice company, for being so patient and always helpful, but mostly for being a good and supportive friend. I also would like to extend my thanks to Anna Odnell. Thank you for always being cheerful, positive and supportive and also for the nice company in the lab, the office and on conferences. To Patrik Nygren, who was present at the very start and close to the end. Thank you for the friendship, for sharing your always (naah, mostly ☺) ingenious thoughts and your sensational sense of humor. Thank you for valuable feedback and comments on this thesis. Lotta Tolstoy Tegler, thank you for all the positive energy, discussions and support throughout the years. Anton Stark, thank you for the pleasant company in the office during the writing of this thesis and for patiently bearing my German swearing whenever Word or Endnote failed to do as I wanted. To Theresa Lindholm, Gabriel Gustafsson, Anna Brynås and Tomas Mårtensson for interesting discussions and nice company.

My gratitude to Susanne Andersson, for nice chats, patiently helping with administrative matters and international shippings. I would like to thank Alexandra Ahlner, for the support and friendship throughout the years and much appreciated feedback on this thesis. Leffe Johansson, you were the first person I met in the lab and we had so much fun ever since. Thank you for always caring, being helpful and supportive and a really good friend-Achtung Bananen, Herr. J! My thanks also to Linda Helmfors, für die Unterstützung, Deinen großartigen Humor, die gemeinsame Zeit und die Freundschaft. To Rozalyn Simon, thank you for interesting and always very inspiring discussions, delicious food and the friendship. To Daniel Kamnert for the nice company, much appreciated help in the lab and the friendship. Furthermore I would like to thank Karin Magnusson for a very enjoyable time as my “office-sambo”, your positive energy and support. Thanks to Janosch Hennig for German beer, late nights out and the friendship.

Additionally, I would like to thank present and former members of the “molecular biotechnology corridor” and the “Torsdags (aka “Nalle”) Fika group” for making the time at work enjoyable, rich in calories and delicious. Not only did I learn how to run exceptional good 2-DE gels, I also learned how to bake gluten-free and egg-free and gained much insight into the Swedish culture and everyday life.
My gratitude to all former and present colleagues and fellow PhD students at IFM-Chemistry for generating an inspiring and supportive work environment. It was great to work at a place where colleagues not only readily shared their experience and knowledge, but also kind words. In this respect my special thanks to Liza, Maria J, Anki, Cissi, Sara, Annica, Alexander, Mathias E, Hanna, Arja, Marcus B, Roger, Per, Peter K, Linda L, Maria L, Patrik L, Vivian, Lasse, Markus N, Peter N, Sofie, Hamid, Maria S, Magdalena, Maria T, Patricia, Veronica, Madhan, Amélie, Daniel S, Ina, Anna-Lena, Therése, Michi, Mathias T, Jeff, Alma, Andreas, Timmy, Jussi and Lan.

My thanks to Stefan Klintström, Charlotte Immerstrand and Anette Andersson at Forum Scientium. Thank you for all the work you put into the graduate school and creating a much appreciated network for education, collaborations and support. I would also like to thank Anke Suska for the friendship and support throughout the years. Thank you for your advice, all the help, nice company, delicious food and great Argentinian desserts.

Life in the academia can be very special and I would like to extend my thanks to my friends and family outside the “academic bubble”. My gratitude to Susanne for your love and support, your great humor and knowing me so well. To Katrin for always being supportive and kind. To Kristina and Dennis, for being the world’s best neighbors and Linda and Wiktor for being such good friends.

My deepest and sincere gratitude to my family for their never-ending love, help and encouragement. I thank my brother Steffen for always being there for me as a friend and being an inspiration at all times. I am forever thankful to my parents Veronika and Michael for giving me the opportunities and experiences that have made me who I am. You are the best parents I could have wished for. I would like to thank my son Tim for making me see the world through different eyes. I am looking forward to being part of your journey.

And finally, to my husband Marcus. You are my very best friend and favorite person. Thank you for your love, support and the countless times you read this thesis. Eine Frage an Dich zum Schluss, Herr Dr. med. Speda: Wer trägt jetzt den Hut der Wissenschaft?
References


Publications

The articles 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-133206