Max-Planck Institute for Marine Microbiology

Department of Molecular Ecology, Bacterioplankton Group & Microbial Genomics and Bioinformatics Group

Master Thesis

(Meta-)genomic Analysis of the Diversity and the Carbohydrate Degradation Potential of the SAR92 Clade during a Diatom-induced Bacterioplankton Bloom

Lennart Kappelmann

15th March 2013

Master thesis submitted to the International Max Planck Research School of Marine Microbiology and the University of Bremen for the partial fulfillment of the requirements to receive the degree

Master of Science
This thesis has been conducted within the Master of Science program of Marine Microbiology (MarMic) at the Max Planck Institute for Marine Microbiology in Bremen, Department of Molecular Ecology of Prof. Dr. R. Amann between September 2012 and March 2013

1st reviewer: Dr. Hanno Teeling

2nd reviewer: PD Dr. Bernhard Fuchs
Declaration Copyright

No data can be taken out of this work without prior approval of the thesis-promoter.

Declaration plagiarism

I hereby confirm that I have independently composed this Master thesis and that no other than the indicated aid and sources have been used. This work has not been presented to any other examination board.

(Ort und Datum / Place and Date)    (Unterschrift / Signature )
# Table of Contents

Summary ................................................................................................................................................ 5
List of abbreviations ................................................................................................................................. 6
1 Introduction ......................................................................................................................................... 7
  1.1 The marine food web ..................................................................................................................... 7
  1.2 The MIMAS project ...................................................................................................................... 9
  1.3 The SAR92 clade ......................................................................................................................... 12
  1.4 Marine carbohydrates .................................................................................................................. 14
  1.5 Carbohydrate-active enzymes (CAZymes) ................................................................................ 15
  1.6 CAZyme analysis ....................................................................................................................... 17
  1.5 Molecular modeling ..................................................................................................................... 18
2 Methods ............................................................................................................................................ 20
  2.1 Phylogenetic analysis of SAR92 ............................................................................................... 20
  2.2 CAZyme annotation ................................................................................................................... 21
  2.3 Molecular modeling using Phyre2 ............................................................................................... 21
3 Results .............................................................................................................................................. 24
  3.1 Phylogenetic analysis of SAR92 ............................................................................................... 24
  3.2 CAZyme analysis ....................................................................................................................... 27
  3.3 Molecular modeling ..................................................................................................................... 30
4 Discussion ........................................................................................................................................ 36
  4.1 Diversity of the SAR92 clade ..................................................................................................... 36
  4.2 Carbohydrate degradation potential ......................................................................................... 37
    4.2.1 CAZyme analysis .................................................................................................................. 37
    4.2.2 Molecular modeling .............................................................................................................. 40
6 Outlook ............................................................................................................................................. 42
8 Acknowledgements .............................................................................................................................. 44
7 References ....................................................................................................................................... 45
**Summary**

SAR92 is a ubiquitously present gammaproteobacterial clade in coastal margin zones. It showed high abundance in the late phase of a diatom-induced bacterioplankton bloom investigated in 2009 in the MIMAS study. The biodiversity and seasonality was investigated at certain dates before and after bloom by 16S rRNA genes derived from clone libraries and metagenomes. SAR92 grouped into three major clusters. Two clusters contained mainly sequences sampled during the bloom while one cluster was mainly comprised of sequences sampled before the bloom.

Analysis of the glycoside hydrolase profile of SAR92 by CAZyme annotation revealed a characteristic pattern of GH families containing enzymes for the synergistic degradation of β-1,3, β-1,4 and β-1,6 linked polysaccharides. Investigation of polysaccharide localization loci of SAR92 strain HTCC2207 and flavobacterial species with similar CAZyme profiles showed similar co-localization of GH16 and TonB-dependent receptors, interacting in the uptake and degradation of specific polysaccharides. On the other hand, SAR92 did not possess SusD binding domains in its genome, indicating potentially distinct substrates.

Subsequent multiple sequence alignment, molecular modeling and in-depth investigation of the active site of a selected metagenomic protein demonstrated strikingly similar topologies of all functionally significant amino acids of the active site and subsites of the enzyme. In all probability SAR92 possesses a laminarinase, enabling SAR92 to grow on the substrates laminarin found in brown algae and chrysolaminarin found in diatoms.
# List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARD-FISH</td>
<td>Catalyzed Reporter Deposition Fluorescence <em>in-situ</em> Hybridization</td>
</tr>
<tr>
<td>CASP</td>
<td>Critical Assessment of protein Structure Prediction</td>
</tr>
<tr>
<td>CAZy (database)</td>
<td>Carbohydrate-active enzyme database</td>
</tr>
<tr>
<td>CAZymes</td>
<td>carbohydrate-active enzymes</td>
</tr>
<tr>
<td>CE</td>
<td>carbohydrate esterase</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DOM</td>
<td>dissolved organic matter</td>
</tr>
<tr>
<td>GH</td>
<td>glycoside hydrolase</td>
</tr>
<tr>
<td>GT</td>
<td>glycosyltransferase</td>
</tr>
<tr>
<td>MIMAS</td>
<td>Microbial Interactions in Marine Systems</td>
</tr>
<tr>
<td>NGS</td>
<td>next-generation sequencing</td>
</tr>
<tr>
<td>NPP</td>
<td>net primary production</td>
</tr>
<tr>
<td>OM</td>
<td>organic matter</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PHYLIP</td>
<td>Phylogeny Inference Package</td>
</tr>
<tr>
<td>PL</td>
<td>polysaccharide lyase</td>
</tr>
<tr>
<td>POM</td>
<td>particulate organic matter</td>
</tr>
<tr>
<td>PUL</td>
<td>polysaccharide localization locus</td>
</tr>
<tr>
<td>RAxML</td>
<td>Randomized Accelerated Maximum Likelihood</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal ribonucleic acid</td>
</tr>
<tr>
<td>SDS</td>
<td>sodiumdodecyl sulfate</td>
</tr>
<tr>
<td>SINA</td>
<td>Silva incremental aligner</td>
</tr>
<tr>
<td>SCOP</td>
<td>Structural Classification of Proteins</td>
</tr>
<tr>
<td>TEP</td>
<td>transparent exopolymer particles</td>
</tr>
<tr>
<td>TBDT</td>
<td>TonB-dependent transporter</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 The marine food web

More than two thirds of our planet’s surface is covered by oceans. They provide the habitat for primary producers, which turn inorganic molecules into organic matter thus forming the basis of life. Marine algae contribute roughly half of the global organic matter net primary production (NPP) [1]. Diatoms are among the most important marine primary producers, being responsible for about one-fifth of the photosynthesis on Earth [2]. Diatom primary production is not constant, but peaks during annually reoccurring blooms which characterize eutrophic upwelling zones and coastal zones in higher latitudes [3, 4]. Although these mostly coastal zones make up only ~7% of the global ocean surface, they account for ~19% of its net primary production (NPP) [5].

Marine primary producers form the basis of a complex food web; they provide a food source for heterotrophic bacteria, especially *Bacteroidetes*, *Alpha* - and *Gammaproteobacteria* [3, 6–8]. These fall prey to protists (flagellates, ciliates) from which the food chain progresses to increasingly larger organisms, e.g. via copepodes to fish. By this mechanism, energy-rich organic carbon compounds are passed on to higher trophic levels.

Conversely, organic matter (OM) is constantly released from this food web by viral lysis, excretion of waste products, exudation of exopolymers (for example by diatoms as transparent exopolymer particles (TEP)) as well as sloppy feeding by zooplankton. OM is divided into particulate organic matter (POM) and dissolved organic matter (DOM). A large portion of the POM, stemming from secreted materials, fecal pellets or debris is decomposed by attaching bacteria [9] or solubilized enzymatically [10] and thereby reintroduced as DOM to the system. Another portion aggregates and sinks to the aphotic zone as marine snow.

DOM forms an immense nutrient reservoir and consists of a broad range of organic molecules usually defined as smaller as 0.45 µm, such as bacterial carbohydrates. 50-100% of the photosynthetically fixed organic matter eventually flows through the DOM pool [11]. Only a small fraction of DOM is readily utilizable by non-bacterial marine organisms at higher trophic levels. This reintroduction of OM to the higher trophic levels of the marine food web is referred to as the microbial loop [12, 13]. It is of utmost importance because it plays a major part in carbon cycling and provides additional energy to the system. Ultimately, it influences issues of global importance like atmospheric CO₂ concentrations through the air sea exchange of carbon dioxide, carbon storage through sinking and carbon flux to fisheries [14].
Since the microbial loop concept was first proposed by Azam et al in 1983, the factors controlling bacterial growth and population size [15, 16] and regulation of its biomass by mortality [17, 18] are investigated intensely for marine ecosystems. Different bacterial taxa have distinct genomic potentials to utilize different types and quantities of DOM. For instance, a radiotracer incubation study by Alonso and Pernthaler [19] showed that members of the alphaproteobacterial SAR11 and Roseobacter clades dominated glucose assimilation at low substrate concentrations, whereas bacteria from the Cytophaga-Flavobacteria lineage dominated at high glucose concentrations. The effects of nutrient availability on the temporal assemblage of bacterioplankton were first studied in monthly sampling studies based on PCR-based methods and revealed that bacterial assemblage patterns reoccurred annually and were affected by both biotic and abiotic factors [20–22].

Considering that bacterial communities can shift within one week or even less, this monthly PCR-based sampling approach is limited in grasping a bacterial succession in a quickly changing environment.
1.2 The MIMAS project

The MIMAS (Microbial Interactions in Marine Systems) BMBF-funded (2008-2011) project focused on the response of North Sea bacterioplankton to spring phytoplankton blooms at high taxonomic and functional resolution. A detailed study of the bacterioplankton during and after a diatom-dominated spring bloom in 2009 revealed a substrate-controlled succession of distinct bacterioplankton clades [3]. Samples were taken in the German Bight near the island of Helgoland, about 50 kilometers offshore at the station “Kabeltonne” at Helgoland Roads, (54°11.03’N, 7°54.00’E, Figure 2), between the main island of Helgoland and a smaller island called “Düne”. “Kabeltonne” is a long-term ecological research site [23], where abiotic factors like temperature and salinity were measured since 1873 and microbiological parameters since 1962. Spring diatom blooms recur annually, providing a suitable area for studying phytoplankton-bacterioplankton interactions in a typical temperate coastal margin zone.

For the MIMAS study, “Kabeltonne” provided a plethora of physicochemical data relevant for the ecological study of bacterial populations in a marine environment. Temperature, salinity, turbidity and the concentrations of silicate, phosphate, nitrate, nitrite and ammonia were measured on a weekday basis. Five hundred liters of subsurface seawater were sampled twice a week and filtered into fractions dominated by particle-associated bacteria (10µm filter...
pore size) and free-living bacteria (0.2µm pore size, GTTP filter). Algal composition was determined microscopically and bacterial *in-situ* abundances were determined via catalyzed reporter deposition fluorescence *in situ* hybridization (CARD-FISH) [25, 26]. At selected sampling times during and after the bloom, the data was complemented by comparative analysis of 16S ribosomal RNA (rRNA) gene amplicons (pyrotags) as well as taxonomic and functional data from DNA (metagenomics) and cDNA (metatranscriptomics) 454 pyrosequencing as well as metaproteome analysis. This approach captured data of the beginning, development and termination of the spring diatom bloom and the subsequent bacterioplankton blooms in 2009.

Pre-bloom bacteria were dominated by oligotrophic *Alphaproteobacteria* of the *Roseobacter* and SAR11 clades, the latter consisting almost entirely of *Candidatus* Pelagibacter ubique.

Chlorophyll a concentration measurements showed that the algae bloom of 2009 (dominated by the diatoms *Thalassiosira* spp. and *Chaetoceros* spp.) started in mid-March, reached its maximum around a week later and gradually abated towards mid-April. The first bacterioplankton response occurred about two weeks after the diatom bloom and was dominated by *Flavobacteria*, whose abundance increased five-fold within one week, especially *Formosa* sp. whose abundance rose from below 1% to 23%. The genera *Reinekea* (class *Gammaproteobacteria*) and *Polaribacter* (class *Flavobacteria*) responded roughly a week later to the diatom bloom. A late increase of abundance was observed for *Gammaproteobacteria* from the SAR92 clade, whose abundance initially increased slightly within two weeks from about 3% to 7% around one week after the diatom bloom started. From April 14th to April 21st, their cell numbers increased 5 fold from ~5 x 10^4 to 2.5 x 10^5 cells/mL, corresponding to a total bacterial abundance of 15%. Within two more weeks, the SAR92 diminished to its pre-bloom abundance.

Two major biological theories can be considered for such a bacterial succession pattern. The first one is “top-down” control, where the bacterial abundance is diminished by predators such as nanoflagellates and viruses who proliferate in response to the plentiful food supply. The second explanation is “bottom-up” control, where species proliferate to a point when their specific substrate is depleted and its population is essentially starving. “Bottom-up” control as well as nutrient niche partitioning provided the best explanations for the findings of the MIMAS study. Metagenomic taxobins of different points in time during the succession showed shifts in gene repertoires in response to a succession of substrates, namely distinct CAZyme and transporter profiles as well as sulfatase numbers.
Figure 3: Absolute cell numbers of selected abundant free-living populations of Alphaproteobacteria [B], Flavobacteria [C] and Gammaproteobacteria [D] during the bacterioplankton bloom as determined by CARD-FISH. Figure adapted from [3].

Metagenomes of 2009 were taken before (Feb. 11th), during (Mar. 31st, Apr 7th, Apr. 14th) and after (June 6th) the phytoplankton bloom, partitioned into taxonomically coherent bins (taxobins) and used for identification, annotation and semi-quantitative analysis of the metaproteome data. This approach allowed the investigation of shifts in gene content and gene expression of dominating bacterial populations.

The MIMAS study found a pronounced peak in the abundance of carbohydrate-active enzymes (section 1.4) during the bacterial succession. Furthermore, TonB-dependent transporter (TBDT) components dominated expressed transport proteins in Flavobacteria and the SAR92 clade [3], which are known to be responsible for the uptake of large compounds, including oligosaccharides.

The results of the MIMAS study suggest that the availability of phytoplankton-derived substrates provided a series of ecological niches in which distinct bacterioplankton
populations could thrive. This means that despite a seemingly homogenous habitat, different populations possess unique genetic potentials and can avoid extinction by direct competition [3].

1.3 The SAR92 clade

SAR92 is a monophyletic clade within the class of Gammaproteobacteria. It is an exclusively marine, heterotrophic group and was discovered as a taxonomically distinct clade during the analysis of environmental gene clone libraries of 16S rRNA genes amplified in 1990 from Sargasso Sea bacterioplankton [27]; hence the SAR abbreviation in the clade’s taxonomic name.

At that point in time, no member of this clade had been isolated yet. The first SAR92 strains were soon isolated from the Oregon coast in 2004 [28] through new high-throughput cultivation methods [29], which allowed large numbers of microbial isolates to be recovered by dilution-to-extinction in natural seawater media [28]. From these strains, Stingl et al chose strain HTCC2207 for research mainly focused on proteorhodopsin, a light-dependent proton pump [30]. This publication from 2007 is the only one specifically concerning SAR92 to date. It led to the first available genome draft of a SAR92 representative, sequenced by the J. Craig Venter Institute as part of the Moore Foundation Microbial Genome Sequencing Initiative. The annotated unclosed genome draft consists of seven contigs with a total length of 2.6 Mbp containing 2,390 open reading frames (ORFs) [30] and is publicly available (GenBank accession number AAPI00000000, version GI:90333150).

The closest cultured relatives of SAR92 so far are found in the genus Microbulbifer, differing more than ~9% from SAR92 in their 16S rRNA genes, indicating that the SAR92 constitute a potential new genus within the Gammaproteobacteria [30]. Furthermore, the SAR92 16S rRNA sequences were found to differ up to 12% among each other, and could be divided into three subclusters. Cultivation of three SAR92 representatives (HTCC2121, HTCC2207 and HTCC2290) from two of these subclusters revealed distinct morphologies (Figure 4) and physiological characteristics such as temperature ranges and optima, maximum cell densities and optimal concentrations of DOC (Figure 5) [31].
FISH and RNA dot blot analyses showed that SAR92 comprised up to 10% of the total bacterial population of the Oregon coast, with decreasing abundance the further the samples got from the coast.

Furthermore, Stingl et al. found that “the peak of abundance correlated with the relatively high nutrient concentrations found in an upwelling region off the Oregon coast. In the lower-nutrient regions farther off the coast, the abundance of the SAR92 was low, close to the limit of detection” [30]. Contrary to those findings, Stingl et al. state that SAR92 is an oligotrophic clade (thriving under low-nutrient concentrations), based on the cultivation success in low-nutrient artificial sea water and the fact that some, but not all of the SAR92 isolates contain proteorhodopsin. This protein is also found in SAR11 representatives like “Candidatus Pelagibacter ubique” and is hypothesized to be able to generate an additional proton motive force without the cost of oxidizing carbon compounds [32]. This would theoretically lead to a more effective use of organic carbon, a desirable feature in nutrient-limited marine environments [33]. However, Stingl et al. could not support this hypothesis, as light did not enhance growth yield or growth rate of HTCC2207 in batch culture. One year later, Stingl et

---

**Figure 4:** Epifluorescence micrographs of DAPI-stained exponential phase of HTCC2121 (J), HTCC2290 (K) and HTCC2207 (L). Scale bars, 1 µm. Taken from [28]

**Figure 5:** Variation of specific growth rates (µ) of SAR92 isolates in the OMG group at different temperatures and DOC concentrations. Taken from [28]
al themselves found that “a growing number of investigations do not readily fit [the assumed benefit of proteorhodopsin for energy gaining] which indicates that proteorhodopsins could have a range of physiological functions” [34]. This now leaves the open question, whether the heterotrophic organisms from the SAR92 clade can be linked to a rather oligotrophic or eutrophic lifestyle.

1.4 Marine carbohydrates

The term carbohydrate is a widely used synonym for saccharides, often also imprecisely referred to as sugars. From their chemical structure, they are accurately polyhydroxy aldehydes and ketones. Carbohydrates consist mainly of carbon, hydrogen and oxygen in a ratio of $C_m(H_2O)_n$ and are occasionally substituted by heteroatoms like nitrogen, phosphorus and sulfur, which is widely available as sulfate at a concentration of 28 mM in marine systems.

Carbohydrates are essential molecules of life. They serve as stabilizing structural compounds in the cell wall matrix and as dense carbon storage compounds. By condensation (the removal of a water molecule), two monosaccharides can be linked via glycosidic bond to form a disaccharide. More precisely, the transfer of the activated sugar moiety from the anomeric hemiacetal center C1 of one monosaccharide to the hydroxylic acceptor of the other monosaccharide. These glycosidic bonds exist in α and β configurations, based on the stereochemical configuration of the anomeric center and the anomeric reference atom of the first monomer. The monomer and the bond are declared α if the two oxygen atoms exist in cis configuration. Vice versa, if the oxygen atoms exhibit trans configuration, the monomer and the bond exist in β configuration. For example, lactose is a disaccharide derived from the condensation of the monosaccharides glucose and galactose, forming a β-1,4-glycosidic bond, indicating that the bond is located between the anomeric C1 carbon atom of the glucose and the C4 atom of the galactose.

Multiple formation of glycosidic bonds between the C1 and possibly any carbon atom of the second monosaccharide lead to the formation of polysaccharides. They can contain branches, which give them a dense structure, hence making them an excellent storage compound. The degree of branching determines whether a polysaccharide is soluble in water or not, with more branching making the polysaccharide less soluble. For example, laminarin is an algal storage compound consisting of β-1,3 linked glucoses occasionally branched with β-1,6 linkages. It’s reducing chain end is infrequently capped with a mannitol molecule.
While terrestrial carbohydrates like glycogen, starch or cellulose are well-studied organic compounds, comparatively little is known about their marine counterparts. They are often anionic due to carboxylic acid residues (e.g. alginate [36]) and sulfate groups. Sulfated polysaccharides include agars, ulvans, fucans, porphyrans and carrageenans [37–39] and require additional sulfatases to be processed. Additionally, their anionic character has been hypothesized to assist in salt-resistance [40, 41].

Carbohydrates are an important carbon source for many marine heterotrophic bacteria. Despite their complex branched and often sulfated character, which requires several enzymes for complete utilization, some organisms have already been shown to possess the enzymatic repertoire to degrade certain carbohydrates, providing the organism with an excellent ecological niche. For example, Hehemann et al found that the marine, agar degrading Bacteroidetes Zobellia galactanivorans DsijT possesses a complex agarolytic system comprising four $\beta$-agarases and five $\beta$-porphyranases, which are able to degrade the sulfated polymers found in various quantities in agars [42].

1.5 Carbohydrate-active enzymes (CAZymes)

CAZymes are found in all domains of life and catalyze reactions of the carbohydrate metabolism of an organism. They are divided into four enzyme classes: glycoside hydrolases (GH) cleave glycosidic bonds by hydrolysis, whereas glycosyltransferases (GT) form glycosidic bonds by the transfer of sugar moieties from activated donor molecules to specific acceptor molecules. Furthermore, polysaccharide lyases (PL) and carbohydrate esterases (CE) are also capable of cleaving the glycosidic bonds.
Information on these highly diverse enzymes is compiled in the CAZy database [43]. CAZymes are classified into families according to sequential and structural similarity, which makes it easy to study their phylogenetic relationships. This structural similarity, for example regarding the catalytic kinetics and mechanisms of an active site, does not specify for only one substrate, as polysaccharides can consist of different monosaccharides, which are linked by the same glycosidic bond [44].

For the analysis of carbohydrate degradation potentials of heterotrophic microbes, glycoside hydrolases are particularly interesting. So far, families GH16, 50, 82, 86, 96, 105, 107, 117 and 118 have been found to degrade marine polysaccharides. The best-characterized among those is GH family 16. It comprises β-1,3, β-1,4 and 1,3-1,4-β-D-glucan endohydrolases, β-1,3 galactanases, β-1,4-galactanases (κ-carrageenases) and xyloglucan transglycosylases/hydrolases [45]. These enzymes can hydrolyze the glycosidic bonds of laminarin (brown algae) [46, 47], chrysolaminarin (diatoms) [48], agar [42], κ-carrageenan [49] and porphyran [37] (all red algae) and lichenin (lichen) [50]. GH16 genes have been found clustered with genes associated with carbohydrate transport [51, 52]. A prominent example of such genes are TonB-dependent receptor genes. TBDTs allow the uptake of large compounds exceeding 600 Da [53] and were found to be highly expressed during bacterioplankton blooms in response to diatom blooms [3]. These clusters are referred to as polysaccharide-utilization loci (PUL) in genomes and are considered operons or regulons.

| Table 1: Confirmed substrates of GH family 16 [37, 42, 45-50] |
|---------------------------------|---------------------------------|
| β-1,3 glycosidic bond           | β-1,4 glycosidic bond           |
| Lichenin                        | Lichenin                        |
| Laminarin                      | Carrageenan                     |
| Chrysolaminarin                 |                                 |
| Agar                            |                                 |
| Porphyran                       |                                 |

Carbohydrate utilization is most efficient when different CAZymes work together synergistically to degrade complexly branched polysaccharides [54, 55]. Cellulose for example is completely degraded by three different enzymes [56, 57]: Endo-acting glucanases (endoglucanases) cleave the β-1,4 glycosidic bonds of long cellulose polysaccharides at random locations. These shortened cellulose chains serve as starting points for exo-acting glucanases (cellbiohydrolases) which accept the pre-processed cellulose via their termini and then cuts into disaccharides (cellubiose). Finally, β-glucosidase
splits the cellubiose into glucose monomers. These three enzymes have a higher combined activity than the sum of their individual activities [54].

The same principle was observed for the degradation of laminarin, a storage glucan found in brown algae. It is a glucan linearly linked by β-1,3 glycosidic bonds with irregular β-1,6 branches occurring in a ratio of 15:1 [48]. Laminarin was found to be degraded by the archaeon *Pyrococcus furiosus* [46], and the gut microbiota of Antarctic krill feeding on diatoms containing chrysolaminarin [58], where endoglucanases, exoglucanases and β-glucosidases were detected. For complete degradation in accordance with the cellulose degradation, endo-acting β-1,3 glucanases from GH16 or GH64, exo-acting β-1,3 glucanases from GH3, GH5 or GH17 as well as β-1,6 glucanases from GH5 or GH30 would provide a perfect CAZyme synergy for complete utilization of laminarin.

1.6 CAZyme analysis

In times of massive amounts of sequence data, BLAST is an effective and accurate heuristic algorithm to identify and annotate CAZyme sequences [59]. Another widespread alignment method is based on profile Hidden Markov Models (HMM) [60], where similar protein sequences are assigned to a protein family with common characteristics. The profile of each family is used to find new family members or to aid further sequence alignment. Profile Hidden Markov modeling captures more sequence information by position-specific scoring, and is able to find more distant homologues than BLAST, however query sequences will only be identified if they belong to known protein families, which can result in poor identification for so far unknown sequences, e.g. from massive environmental sequencing. HMMs are implemented in the software package HMMER [61], and protein families are collected in the Pfam database [62].

BLAST and HMMER were used to align CAZymes in genomes and metagenomes [63–69]. Results were filtered with stringent cutoff settings by E-value and bit score. The bit score gives an indication of how good the alignment is, taking into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences [70]. The expect value (E) is a parameter describing the possible number of hits by chance. The lower the E-value, the more “significant” the match is. The E-value can therefore be used as a convenient way to set a significance threshold for reporting results.

Since next-generation sequencing technologies like 454-pyrosequencing or Illumina generate sequence data in the gigabase range per run containing potentially thousands of CAZymes and metagenomic environmental studies aim at capturing a big ecological picture,
manual annotation is neither necessary nor practical; in fact automated annotation is the only effective way to do it.

A CAZy profile of an organism, depicting the simple repertoire of CAZymes, can already reveal key characteristics about its lifestyle. An organism living in a eutrophic environment will usually possess a wider range of GH families than an organism living in an oligotrophic environment. The MIMAS project (section 1.2) highlighted the strength of CAZyme profiling in environmental metagenomic studies by identifying lifestyles of key players given enough sequencing coverage & supporting data like key player genomes.

Going into detail, CAZyme genes of interest can be chosen and manually studied further. An alleged CAZyme (proposed by the CAZy annotation pipeline) can be validated by checking the consensus pattern of its active site. These active site patterns can for example be found in the PROSITE database.

By performing an alignment of the translated gene of interest and a number of known representatives of this protein family (e.g. GH family) and subsequent tree construction, closely related proteins with specific substrates can be found and conclusions on structural and functional relationships between these proteins can be drawn.

1.7 Molecular modeling

It is essentially possible to get from sequence affiliation to molecular structure and function by computational modeling. The structure of a protein can be predicted by comparison with sequences of known 3D structure. Due to exponentially growing databases and ever-improving matching algorithms, so called template-based homology modeling or fold-recognition techniques are a quickly emerging approach for obtaining molecular maps of proteins without elaborate, cost intensive laboratory techniques like X-ray crystallography.

Homology modeling methods rely on the observation that the number of folds in nature appears to be limited and that remotely homologous protein sequences adopt remarkably similar structures [71]. In other words, these methods consider the fact that a protein’s secondary structure is evolutionary more conserved than its primary amid acid sequence. Profile HMMs and, most recently, profile–profile matching algorithms [72] are powerful and accurate in aligning remotely related sequences because they capture the mutational propensity of each position in an amino acid sequence based on observed mutations in related sequences.
Using these methods, is it nowadays possible to accurately detect and model protein sequences with less than 20% sequence identity to a known protein [73]. Nevertheless, results require human expertise in analyzing the results in the context of biological knowledge. Practical applications of protein structure prediction are manifold, from developing and assisting the development of functional hypotheses on enzymes [74], improving phasing signals in crystallography [75], selecting sites for mutagenesis [76] to the design of pharmaceutical drugs [77].

1.8 Aims of this Thesis

The SAR92 clade showed high abundances of up to 15% after the spring phytoplankton bloom in the MIMAS project [3] and up to 10% in other coastal margin zone studies [27–29]. Nevertheless, only a single study has been dedicated to the SAR92 clade yet, focusing mostly on proteorhodopsin [30]. This Master Thesis seeks to shed more light on the diversity and carbohydrate degradation potential of this diverse group of marine bacteria. For this purpose, a large variety of bioinformatic methods were applied and a molecular modeling technique was implemented for the first time for in-depth analysis of annotated CAZymes.
2 Methods

2.1 Phylogenetic analysis of SAR92

16S rRNA sequences retrieved during the MIMAS studies, consisting of full-length sequences from clone libraries as well as metagenomically derived sequences were used for the phylogenetic analysis of the SAR92 clade.

16S rRNA sequences from clone libraries were generated by Mariette Kassabgy from the Molecular Ecology department of the Max-Planck-Institute for Marine Microbiology in Bremen. Clone libraries were obtained by initial DNA extraction using the SDS-based DNA extraction method according to Zhou [78], followed by PCR amplification of the 16S rRNA genes. Cleaned PCR products were cloned into pCR4 TOPO vector (Invitrogen, Groningen, Netherlands) which were subsequently transformed into competent E. Coli cells (Invitrogen, Groningen, Netherlands) by heat shock and plated on selective agar for blue-white screening. Clones with successful inserts were Sanger sequenced (GATC Biotech AG, Konstanz, Germany). Sequences were automatically assembled and quality checked using the DNA base sequence assembly software (Heracle BioSoft S.R.L., Pitesti, Romania).

Metagenomic 16S rRNA sequences were extracted from the unassembled reads using the NGS-pipeline implemented in the SILVA project (release 111) [79]. The reads were preprocessed by quality control and alignment: Reads shorter than 150 bp or with more than 2% of ambiguities or more than 2% of homopolymers were removed and the remaining reads were aligned. Unaligned reads were not considered in downstream analysis to eliminate non-16S rRNA gene fragments. The raw data of eight MIMAS metagenomes are publicly available from the European Bioinformatics Institute ENA archive (study number ERP001227)

Phylogenetic analyses were performed using the ARB software package (version 5.5) [80] and the 16S rRNA database SSURef version 111 [81]. Sequences were aligned using the Silva Incremental Aligner (SINA) [82] as implemented in ARB, and manually inspected and improved considering the 16S rRNA molecule’s secondary structure. The full length sequences from the clone libraries were added using the positional variability filter, which defines the probability of observing a given base at a given alignment position. It is calculated based on all bacterial 16S rRNA sequences in the SILVA database with a parsimony approach. For the metagenomic sequences, a 50% positional variability filter was applied, eliminating poorly conserved positions. The metagenomic sequences required an
additional termini filter specifying the boundaries of the 16S rRNA gene because they can contain flanking sequences from other genes which need to be excluded for tree calculation.

The trees were calculated using distance matrix (ARB neighbour joining), maximum parsimony (Phylip DNAPARS) [83] and maximum likelihood (RAxML) [84] methods. A and consensus trees was built from all three methods.

2.2 CAZyme annotation

For CAZy annotation, a pipeline called Trident (developed by Sixing Huang of the Microbial Genomics and Bioinformatics group of the Max Planck Institute for Marine Microbiology in Bremen) was used, which combines three automated annotation approaches and computes a consensus annotation for each sequence. By considering advantages and disadvantages of each of the three approaches, it proved to find a good balance of quickness, sensitivity and specificity.

The pipeline consists of three annotation tools and a tool building the consensus output. The first annotation tool uses a BLAST [59] search against the CAZyme sequences, which were previously collected from the CAZy website [43] and formatted for BLAST. Hits are only kept for E-values lower than E-30. The second annotation tool uses a HMMER search against the Pfam database, where results are semantically linked to CAZy by rules defined by CAZy and Park et al [65]. The third annotation tool uses a HMMER search against the dbCAN database [69]. Both HMMER annotation tools were set to an E-value cutoff of E-15. The consensus-building algorithm then applies the best-hit or the majority-decision logic to provide the final result. Benchmarks were performed to calibrate and adjust by pipeline by predicting CAZymes of public genomes and comparing the results to those provided on the CAZy website. The pipeline generally predicts slightly more CAZymes than reported by the CAZy database, which result from multi-domain CAZymes. Also, short sequences can be incorrectly assigned to a certain protein family when it lacks the catalytic site which defines this family but otherwise aligns well with a reference sequence from that protein family. On the other hand, unknown CAZymes cannot be annotated and lead to ‘underestimation’ of the CAZyme repertoire in that aspect.

2.3 Molecular modeling using Phyre2

Sequences annotated as CAZymes were checked for presence of the consensus pattern of their respective family. Subsequently, these protein sequences were aligned using the
multiple sequence alignment tool of the software package HMMER 3.0, which aligns homologous sequences using profile-HMMs [61, 85]. Protein trees were calculated using Fasttree [86] to infer phylogenies of the alignments.

The molecular structure of selected proteins from these protein trees were then predicted by the web-based tool Phyre2 [73]. This tool contains a database of currently over 65,000 sequences of known 3D structure, coming from the Structural Classification of Proteins (SCOP) database [87], augmented with newer depositions from the Protein Data Bank (PDB) [88]. Each of these structurally known sequences has been PSI-BLASTed against a nonredundant sequence database and a profile HMM has been constructed and stored in a ‘fold library’, together with the known and predicted secondary structure of these proteins.

A submitted query sequence is scanned against the nonredundant sequence database and a profile is constructed. The secondary structure is then predicted using three independent prediction tools: Psi-Pred [89], JNet [90] and SSPro [91]. Each program predicts alpha helices, beta strands and coils in a three-state prediction and provides confidence values at each amino acid position. These confidence values are averaged and a final consensus prediction is created.

The profile and the predicted structure of the query sequence is then scanned against the database of structurally known sequences using a profile-profile alignment algorithm [92]. The top 10 alignments are used to construct the 3D model of the query, giving spatial coordinates for each amino acid of the protein. Confidence and coverage values are provided to evaluate the quality of the molecular prediction in the search output result.

The prediction model of Phyre2 has been shown to be very reliable and powerful. Phyre/Phyre2 frequently ranks amongst the top performing servers in the CASP (Critical Assessment of protein Structure Prediction) international blind trials of structure prediction in homology modeling and remote fold recognition. A more detailed description of all methods used in Phyre2 can be found in Bennett-Lovsey et al [92].

The 3D prediction of the protein was subsequently investigated using the molecule imaging software package UCSF Chimera version 1.7 [93].
Figure 7: Phyre2 prediction model (modified from the Phyre2 website, section help/teaching slides)
3 Results

3.1 Phylogenetic analysis of SAR92

A phylogenetic consensus tree was built from 66 full-length 16S rRNA sequences from the SAR92 clade (Figure 8). The sequences stemmed from DNA extractions from February 11\(^{th}\) and April 14\(^{th}\) of 2009 and ranged between 1484 and 1493 bases.

The 16S rRNA genes could be divided into three larger sub-clusters. Cluster A consisted of 28 sequences exclusively from April and had two smaller branches in close phylogenetic distance. Cluster B contained 24 sequences, 23 of which came from clones from February, whereas only one sequence stemmed from sampling in April. Cluster C formed a smaller group of seven sequences from April which are more distantly related to the other clusters. As outgroup, 16S rRNA genes of *Saccharophagus degradans*, *Microbulbifer hydrolyticus* and *Marinobacter maritimus* were chosen.

![Figure 8: Phylogenetic relationship of the SAR92 clade based on clonal 16S rRNA gene sequences. Consensus tree based on distance matrix (ARB neighbour joining), maximum parsimony (Phylip DNAPARS) and maximum likelihood (RAxML) methods.](image)

Distance matrix revealed high average identities of at least 99% within the three subclusters of the clonal sequences (cluster A: 99.3%; B: 99.5%; C: 99.0%). The average identity between the distinct subclusters of ranged from 96.9% to 97.6%.
Table 2: Average identity between distinct clusters of SAR92 clones

<table>
<thead>
<tr>
<th></th>
<th>Cluster A</th>
<th>Cluster B</th>
<th>Cluster C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster A</td>
<td>-</td>
<td>97.6%</td>
<td>96.9%</td>
</tr>
<tr>
<td>Cluster B</td>
<td>97.6%</td>
<td>-</td>
<td>96.9%</td>
</tr>
<tr>
<td>Cluster C</td>
<td>96.9%</td>
<td>96.9%</td>
<td>-</td>
</tr>
</tbody>
</table>

To investigate whether the metagenomic 16S rRNA sequences generated during the MIMAS study complement the pattern of the three clonal subclusters, 250 16S rRNA sequences from the established SAR92 taxobins were imported and aligned in the existing tree. The metagenomic 16S rRNA sequences were retrieved in 2009 on February 11th, March 31st, April 7th, April 14th and June 6th. Only metagenomic sequences longer than 150 bases were selected. The average read length was 380 bp, ranging from 161 to 560 bp.

Additionally, 16S rRNA sequences of SAR92 strains cultured by Cho and Giovannoni [31] using high-throughput cultivation (code HTCC) were added to the tree. They included strain HTCC2207 with a published genome draft (NCBI Tax ID: 314287).

Table 3: Overview of the metagenomic 16S rRNA sequences assigned to SAR92 taxobins from the MIMAS study

<table>
<thead>
<tr>
<th>Situation</th>
<th>Date</th>
<th>No. of seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-bloom</td>
<td>02/11/2009</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>03/31/2009</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>04/07/2009</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>04/14/2009</td>
<td>63</td>
</tr>
<tr>
<td>post-bloom</td>
<td>06/06/2009</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>06/16/2009</td>
<td>37</td>
</tr>
<tr>
<td>Σ</td>
<td>250</td>
<td></td>
</tr>
</tbody>
</table>

The metagenomic 16S rRNA sequences generally grouped well into the subclusters according to their respective sampling dates. Almost half of them (115 sequences) grouped closely to cluster A and the two smaller neighboring clusters, all containing clonal sequences exclusively from April (Figure 9). Therefore, cluster A was extended to contain these two small subclusters as well for a better overview for the tree containing the metagenomic 16S rRNA sequences. 99 of the 115 sequences (86%) were sampled in the time span from March 31st to April 14th, where SAR92’s abundance increased after the diatom bloom. Almost half of the metagenomic 16S rRNA sequences from February also group together with clonal sequences from February in cluster B. Finally, 56 of the 79 metagenomic 16S rRNA
sequences (70%) which grouped into cluster C were sampled in the time span from March 31st to April 14th. Furthermore, the metagenomic 16S rRNA sequences formed two additional subclusters, namely a post-bloom cluster with 5 sequences from June 16th and a cluster more distantly related to all others from April 14th.

All HTCC reference 16S rRNA sequences fall within the different SAR92 subclusters of the 16S rRNA sequences derived from the MIMAS study. The study of Stingl and coworkers [30] on SAR92 actually showed clustering into 3 subclusters as well, which are reproduced in this phylogenetic tree. HTCC2207 groups closely with subcluster C consisting mainly of sequences sampled during the bloom. Only those HTCC sequences which aligned slightly outside of the clusters can be seen in figure 9, the others are contained within them.

Figure 9: Phylogenetic relationship of the SAR92 clade based on metagenomic, clonal and reference 16S rRNA gene sequences. Consensus tree based on distance matrix (ARB neighbour joining), maximum parsimony (Phylip DNAPARS) and maximum likelihood (RAxML) methods.
3.2 CAZyme analysis

For the CAZyme annotation pipeline reported 24 GH families for the SAR92 strain HTCC2207. 19 GH families for reported for the SAR92 taxobin of April 14th, where CARD-FISH results from the MIMAS study reported the highest abundance (7%) of SAR92 among the available metagenomes [3]. Since this study focuses on the polysaccharide degradation capacities of SAR92, CAZymes involved in the formation of polysaccharides will be neglected and the following results will focus on interpretation of the GH families.

An interesting and well described GH family with a high number of predicted ORFs from this CAZyme profile is GH16. It is known that this family hydrolyzes $\beta$-1,3 and $\beta$-1,4 glycosidic bonds of marine polysaccharides via endo-acting enzymes [94, 95]. Another interesting family is GH17, which has very similar capacities as GH16 but whose enzymes were found to have exo-enzymatic activity [96]. GH5 and GH30 are capable of hydrolyzing $\beta$-1,6 glycosidic bonds, which in collaboration with GH16 and GH17 lead to the synergistic capacity to completely degrade algal polysaccharides with a $\beta$-1,3, $\beta$-1,6 branching pattern like laminarin or chrysolaminarin. The SAR92 taxobins contained three families which were not annotated in the draft genome of HTCC2207, namely GH5, GH31 and GH117.

<table>
<thead>
<tr>
<th>GH Families</th>
<th>HTCC2207</th>
<th>SAR92 taxobin</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH3</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>GH5</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>GH10</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>GH13</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>GH15</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH16</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>GH17</td>
<td>3</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GH Families</th>
<th>HTCC2207</th>
<th>SAR92 taxobin</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH18</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH23</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>GH28</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>GH30</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>GH31</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>GH36</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>GH42</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>GH43</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GH47</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GH Families</th>
<th>HTCC2207</th>
<th>SAR92 taxobin</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH71</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH72</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>GH73</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>GH81</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>GH84</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH95</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH99</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH103</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH117</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 10: GH profile of SAR92 strain HTCC2207 and the SAR92 taxobin from April 14th. The number of predicted ORFs is given below the GH family.
The GH profile of SAR92 strain HTCC2207 was compared with the 20 most closely related organisms available in the CAZy database. Only marine gammaproteobacterial strains with a heterotrophic lifestyle were selected. This approach grants a comparative impression of the diversity of degradable carbohydrates of this SAR92 strain.

The number of GH families of the closely related species varied greatly, ranging from 5 GHs in Oceanimonas sp. GK1 and Alcanivorax borkumensis to 44 GHs in Saccharophagus degradans. HTCC2207 contained 24 GH families on its draft genome of 2.62 Mb, which equals ~9.1 GHs/Mb. In comparison, the closely related marine Gammaproteobacteria contained only an average of 15.4 GH families on an average genome length of ~4.5 Mb, corresponding to only 3.2 GHs/Mb.

Table 4: Comparison of CAZyme profiles of SAR92 strain HTCC2207 and closely related marine Gammaproteobacteria

<table>
<thead>
<tr>
<th>species</th>
<th>order</th>
<th>Genome size (Mb)</th>
<th>No. of GH families</th>
<th>No. Of GH per megabase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marinobacter hydrocarbonoclasticus ATCC 49840</td>
<td>Alteromonadales</td>
<td>3.99</td>
<td>6</td>
<td>1.50</td>
</tr>
<tr>
<td>Glaciecola nitratireducens FR1064</td>
<td>Alteromonadales</td>
<td>4.13</td>
<td>11</td>
<td>2.66</td>
</tr>
<tr>
<td>Teredinibacter turnerae T7901</td>
<td>Alteromonadales</td>
<td>5.19</td>
<td>42</td>
<td>8.09</td>
</tr>
<tr>
<td>Simiduia agarivorans</td>
<td>Alteromonadales</td>
<td>4.31</td>
<td>24</td>
<td>5.57</td>
</tr>
<tr>
<td>Oceanimonas sp. GK1</td>
<td>Alteromonadales</td>
<td>3.51</td>
<td>5</td>
<td>1.42</td>
</tr>
<tr>
<td>Saccharophagus degradans</td>
<td>Alteromonadales</td>
<td>5.06</td>
<td>44</td>
<td>8.70</td>
</tr>
<tr>
<td>Pseudoalteromonas atlantica</td>
<td>Alteromonadales</td>
<td>5.19</td>
<td>7</td>
<td>1.35</td>
</tr>
<tr>
<td>Ferrimonas balearica DSM 9799</td>
<td>Alteromonadales</td>
<td>4.28</td>
<td>8</td>
<td>1.87</td>
</tr>
<tr>
<td>Shewanella denitrificans OS217</td>
<td>Alteromonadales</td>
<td>4.55</td>
<td>13</td>
<td>2.86</td>
</tr>
<tr>
<td>Psychromonas ingrahamii 37</td>
<td>Alteromonadales</td>
<td>4.56</td>
<td>14</td>
<td>3.07</td>
</tr>
<tr>
<td>Idiomarina loihiensis L2TR</td>
<td>Alteromonadales</td>
<td>2.84</td>
<td>6</td>
<td>2.11</td>
</tr>
<tr>
<td>Colwellia psychrerythraea 34H BAA-681</td>
<td>Alteromonadales</td>
<td>5.37</td>
<td>16</td>
<td>2.98</td>
</tr>
<tr>
<td>Marinomonas mediterranea MMB-1</td>
<td>Oceanospirillales</td>
<td>4.68</td>
<td>13</td>
<td>2.78</td>
</tr>
<tr>
<td>Halohemila chejuensis</td>
<td>Oceanospirillales</td>
<td>7.22</td>
<td>18</td>
<td>2.49</td>
</tr>
<tr>
<td>Alcanivorax borkumensis</td>
<td>Oceanospirillales</td>
<td>3.12</td>
<td>5</td>
<td>1.60</td>
</tr>
<tr>
<td>Halomonas elongata DSM 2581</td>
<td>Oceanospirillales</td>
<td>4.06</td>
<td>11</td>
<td>2.71</td>
</tr>
<tr>
<td>Chromohalobacter saltexigens DSM 3043</td>
<td>Oceanospirillales</td>
<td>3.7</td>
<td>9</td>
<td>2.43</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa PAO1</td>
<td>Pseudomonadales</td>
<td>6.26</td>
<td>13</td>
<td>2.08</td>
</tr>
<tr>
<td>Acinetobacter calcoaceticus PHEA-2</td>
<td>Pseudomonadales</td>
<td>3.86</td>
<td>6</td>
<td>1.55</td>
</tr>
<tr>
<td>Aeromonas hydrophila ATCC 7966</td>
<td>Aeromonadales</td>
<td>4.7</td>
<td>19</td>
<td>4.04</td>
</tr>
<tr>
<td><strong>average of closely related species</strong></td>
<td></td>
<td><strong>4.53</strong></td>
<td><strong>14.5</strong></td>
<td><strong>3.20</strong></td>
</tr>
<tr>
<td><strong>SAR92 strain HTCC2207</strong></td>
<td></td>
<td><strong>2.62</strong></td>
<td><strong>24</strong></td>
<td><strong>9.16</strong></td>
</tr>
</tbody>
</table>

The SAR92 taxobin was found to display similar CAZyme and transporter expression profiles as the taxobins of Flavobacteria. Both taxonomic clades showed high abundances of GH16.
and TBDT components. For that reason, GH16 containing polysaccharide localization loci (PULs) of flavobacterial representatives *Zobellia galactanivorans* DsiJT [37], *Gramella forsetii* KT0803 [97], *Flavobacterium johnsoniae* UW101 [98] and *Cellulophaga Lytica* DSM 7480 [99] were compared with a GH16 containing PUL from SAR92 strain HTCC2207.

GH16 and TBDTs were found to be co-localized in PULs of both the flavobacterial species and the SAR92 strain HTCC2207. The PULs of the flavobacterial species showed similar gene patterns which all contained GH16, TBDT and SusD coding genes. *Z. galactanivorans* DsiJT exhibited a co-localized sulfatase while *G. forsetii* KT0803 and *C. Lytica* DSM 7480 displayed additional GH coding genes in their respective PULs. SusD plays a role in substrate binding of the polysaccharide on the cell surface [53, 100–102] and was not present in the HTCC2207 genome. In contrast, it exhibited genes coding for GH17, Plug and ABC transport. Plug was also found to play a role in the TonB system [103, 104] while ABC transporters are involved in a plethora of translocation processes across cellular membranes [105] and could possibly play a role in mono- and disaccharide transport as well.

**Flavobacteri**

<table>
<thead>
<tr>
<th>Flavobacterium</th>
<th>GH 16</th>
<th>TBDT</th>
<th>SusD</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Z. galactanivorans</em> DsiJT</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><em>G. forsetii</em> KT0803</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><em>F. johnsoniae</em> UW101</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td><em>C. lytica</em> DSM 7480</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

**SAR92**

<table>
<thead>
<tr>
<th>SAR92</th>
<th>GH 16</th>
<th>Plug</th>
<th>ABC Transporter</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTCC2207</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

**Figure 11:** Comparison of polysaccharide utilization loci (PULs) of four flavobacterial species and SAR92 strain HTCC2207, containing genes involved in polysaccharide uptake and
degradation. GHs were annotated based on the CAZy database, while all other genes were annotated based on Pfam.

3.3 Molecular modeling

From the GH families reported for the SAR92 taxobin, GH16 was chosen for further in-depth analysis. GH16 is a well characterized group with a conserved, known active site consensus pattern in the PROSITE database. Furthermore, it was present in the SAR92 taxobin with high gene numbers (ORFs) and the overall GH pattern of SAR92 already suggested the synergistic degradation capacity of polysaccharides with a β-1,3, β-1,6 branching pattern as present in laminarin or chrysolaminarin (section 3.2), both degraded by the enzyme laminarinase. Laminarinases are known to be contained in GH16 [94].

For these reasons, a protein tree was calculated by multiple sequence alignment for CAZyme family GH16. It contained amino acid sequences of the SAR92 taxobin derived from the metagenome from April 14\textsuperscript{th} (with the highest SAR92 abundance), from the SAR92 strain HTCC2207 and from known GH16 sequences of different specific enzymes. The sequences were manually checked to contain the active site pattern of GH16. It is important to note that the sequences derived from metagenomic SAR92 taxobins do not necessarily contain the complete protein from N-terminus to C-terminus. The protein tree allows to investigate which characterized enzymes group closest with the SAR92 proteins from metagenomic taxobins and strain HTCC2207 (Figure 12).

Substrate specificity is well represented in the GH16 tree, as laminarinases, κ-carrageenases, licheninases, porphyranases and agarases all group together in distinct clades. SAR92 proteins from metagenomic taxobins and strain HTCC2207 all fell within a laminarinase cluster in the calculated tree.
Figure 12: GH16 Protein tree containing sequences from a metagenomic SAR92 taxobin, SAR92 strain HTCC2207 (both bold) and selected characterized proteins from the GH16 family.

The metagenomic GH16 sequences from the laminarinase cluster were subsequently submitted to the Phyre2 protein fold prediction service and the result with the highest score was selected for detailed investigation of its 3D structure. For Ref. ID 676545, Phyre2 reported a confidence of 100% in the prediction and a coverage of 90%, meaning that the top scoring template used for prediction covered 90% of 676545’s amino acid sequence, a laminarinase from *Thermotoga maritima* (PDB ID: 3B01) henceforth referred to as reference laminarinase.

The topology of the putative laminarinase was investigated by modeling its electrostatic surface. The active site topology of glycoside hydrolases can provide valuable information
about the processing of the substrate. Three general classes exist, which can, in principle, be built on the same fold and the same catalytic residues:

Figure 13: Active site topologies found in glycoside hydrolases. (a) pocket, (b) cleft and (c) tunnel. The catalytic residues are shaded in red. Modified from [106].

The “pocket” or “crater” structure can only accept a substrate via its terminus. Enzymes with this topology are therefore called exo-acting enzymes or exoenzymes. The “cleft” or “groove” structure allows random binding of linear polymeric substrates and can be found in endo-acting enzymes. They interact once, then disengage and engage another polysaccharide strand. Finally, a “tunnel” structure is possible when the protein evolves long loops that narrow the cleft of an endo-acting enzyme. The substrate can be threaded through this tunnel, allowing a progressive processing of the substrate. This means that sugar monomers can be released after hydrolysis while the glycoside hydrolase remains bound to the polysaccharide chain and continues to interact [106].

The putative laminarinase revealed a narrow C-shaped topology, indicating an endo-acting, possibly processive activity.

Figure 14: Putative laminarinase (Reg. ID 676545) depicted as balls and sticks molecule pattern, ribbon style and electrostatic surface style.

The putative SAR92 laminarinase was superimposed with the reference laminarinase, a tetramer of homologous subunits all containing the active site pattern of GH16. The active
sites face the outside of the oligomer so that substrates will likely come in contact with the active site, avoiding blind spots. The superposition also showed that not the whole laminarinase monomer is covered by the metagenomic SAR92 laminarinase, spanning 176 amino acids as compared to the 258 amino acids of a complete monomer from the reference laminarinase.

Figure 15: Superimposed putative SAR92 laminarinase and laminarinase tetramer from *Thermotoga maritima* (PDB ID: 3B01). The cutout on the right shows the largely overlaying ribbons of both enzymes in the active site region. On the left, the putative SAR92 laminarinase is depicted by its molecular surface to illustrate the size proportions relative to the whole reference tetramer.

Figure 16: Superimposed active site of putative SAR92 laminarinase and the reference laminarinase from *Thermotoga maritima* (PDB ID: 3B01).
The active site of the putative SAR92 laminarinase was subsequently inspected in further
detail. Its folding topology consisted of two antiparallel β-sheets that stack to form a β-
sandwich consisting of one concave and one convex face. The concave face, meaning that
the amino acid residues face the inside of the curved β-sheet, contains the sugar substrate
binding sites, with the active site pattern located on a single strand that is offset from the
center of the sheet. This structure was already described by Ståhlberg et al. in 1996 [107]. All
GH16 sequences have an active site consensus pattern of ExDxxE. Considering all
confirmed amino acids found in this GH16 consensus pattern so far, it can be specified as E-
[LIV]-D-[LIVF]-x(0,1)-E-x(2)-[GQ] according to PROSITE.

Among the seven known laminarinase representatives from figure 12, six showed an EIDIME
motif. The putative SAR92 laminarinase had the pattern EIDMME, meaning that the unpolar
isoleucine (I) was substituted by an unpolar methionine (M). The most important amino acids
in this motif are however the two glutamic acids enclosing it. The first glutamyl residue has
been unambiguously identified as the catalytic nucleophile, while the second glutamyl
residue serves as the general acid/base. The aspartate (D) in the middle has been described
to form a tight hydrogen bond with the nucleophile [107–109].

The reaction that takes place in GH16 has been described in detail in “Carbohydrates in
Chemistry and Biology”, Chapter 29: “Glycosidase-Catalysed Oligosaccharide Synthesis”
[95]. The first side chain carboxylate of the glutamyl residue of the glutamic acid acts as a
nucleophile and attacks the anomeric carbon of the sugar ring. A proton from a second
carboxylate facilitates the fission of the glycosidic bond by general acid catalysis. The
glycosyl enzyme intermediate thus formed is decomposed in a second step, wherein a water
molecule is activated, leading to a nucleophilic attack on the anomeric center by general
base catalysis from the deprotonated form of the second carboxylate [110]. This second
carboxylate is commonly referred to as the general acid/base residue because of its dual role
in catalysis. [94]

In the iterative hydrolysis of long polysaccharides, the locations of the sugar units are
numbered upstream and downstream of the reaction site, according to a nomenclature
introduced by Davis et al [111]. Subsite -1 is one position before the reaction site and subsite
+1 is one position after the reaction site. In both the putative SAR92 laminarinase and the
reference laminarinase, three conserved tryptophans can be found near the active site. Like
in many other sugar binding enzymes, aromatic residues like the planar indole residues of
the tryptophans can establish van der Waals interactions with potential polysaccharide
substrates [109], thus coordinating the sugar units from upstream subsites to the active site,
holding them in place during hydrolysis, and coordinating them out again.
Furthermore, the putative laminarinase contains an arginine pointing towards the active site. Arginine has been described to be assist in the substrate distortion by binding the sugar in the +1 subsite. The distortion creates space in the active site which enables positioning so that the general acid/base residue can interact with the glycosidic oxygen.[112]

Figure 17: Active site similarities of two GH16 glucanases. Left: GH16 glucanase from *Phanerochaete chrysosporium* with a disaccharide as substrate (modeled by Aurore Labourel from Station Biologique Roscoff) Right: Active site of the putative SAR92 laminarinase with laminaribiose as substrate. The residues of amino acids involved in the reaction (glutamate, arginine) and holding the substrate in place (tryptophan) are shown as balls and sticks patterns.
4 Discussion

4.1 Diversity of the SAR92 clade

The phylogenetic analysis revealed a distinct clustering within the SAR92 clade. Three major clusters formed which were dominated by sequences from specific sampling dates. Two large clusters (A and C) contained a majority of sequences (86 and 70%) sampled during the SAR92 bloom after the diatom bloom. Their full length 16S rRNA sequences had an average identity of 96.9%, thus potentially representing distinct genera within the SAR92 clade. The third cluster (B) was to a large extent comprised of sequences sampled in winter before the diatom bloom. This pre-bloom cluster showed average identities of 97.6% and 96.9% with the previously mentioned bloom clusters.

The finding of three large clusters within the SAR92 clade confirms the result of the phylogenetic analysis proposed by Stingl et al in 2007 [30]. In fact, their HTCC sequences arranged into the three clusters of this tree in an identical way as they did in the Stingl. This is remarkable considering the fact that the sequences from the MIMAS study were sampled in the North Sea and the HTCC sequences were sampled at the Oregon coast of the Pacific. The North Sea constantly receives inflow of North Atlantic sea water which is coupled to the global conveyor belt, which eventually flows through the Pacific. Since SAR92 sequences from two that distantly related sampling sites group in such a similar pattern, it is possible that the diversity of the SAR92 clade with respect to this 3-cluster-appearance has been well assessed.

It is of course very likely that further clusters within the SAR92 exist. This is indicated by the two additional metagenomic clusters, which revealed another more distantly related cluster of SAR92 sampled during the bloom and, interestingly, a post-bloom cluster of metagenomic sequences sampled in the summer.

Considering the overall phylogenetic pattern revealed in this analysis, it seems possible that different subgroups within SAR92 could to some extent possess different lifestyles. It would be reasonable to assume that the pre-bloom cluster thrives under more nutrient-limited conditions than the clusters found after the diatom bloom. Stingl et al. found that only some of their SAR92 isolates possessed a proteorhodopsin gene, and all of those fall within cluster C sampled after the diatom bloom. A possible explanation is that SAR92 representatives of this cluster are able to live under nutrient-limited conditions through proteorhodopsin (aiding in generating a proton-motive force) and are opportunistic when a degradable substrate is provided by the diatom bloom.
SAR92 was constantly present at low abundance in the winter of 2009 (before the diatom bloom). The substrate during those times can of course only be speculated. Brown algae from the Laminaria genus are known to grow year-round even in arctic waters [113], providing a potential substrate in laminarin. When the diatoms were disrupted after their bloom, their carbon storage polysaccharide chrysolaminarin was released, thus providing an additional substrate for SAR92. It could of course also be a completely different substrate unassessed in this study, as not all GH families could be investigated in sufficient detail.

4.2 Carbohydrate degradation potential

4.2.1 CAZyme analysis

The CAZyme profiles of HTCC2207 and the SAR92 taxobin from April 14th display a relatively broad spectrum of 24 glycoside hydrolases for HTCC2207 and 19 in the SAR92 taxobin. CAZyme profiles have to be evaluated in context, and the profile of HTCC2207 can only provide a first hint at its actual lifestyle. In comparison, “Candidatus Pelagibacter ubique” HTCC1062, a typical oligotrophic organism from the SAR11 clade, has merely a total of four GH families. On the other hand, the Flavobacterium Zobellia galactanivorans DsiJT possesses a total of 40 GH families, showing that also much higher GH diversities are possible.

Comparison of the CAZyme profile of SAR92 strain HTCC2207 with closely related, marine gammaproteobacterial strains with a heterotrophic lifestyle revealed ~9.1 GHs per megabase in the SAR92 strain compared to ~3.2 GHs per megabase in the closely related Gammaproteobacteria, corresponding to around 2.8 times more GHs per megabase in the SAR92 draft genome.

The CAZyme profiles of SAR92 therefore do not indicate a typical oligotrophic lifestyle, as claimed by Stingl and coworkers [30], but rather a eutrophic, possibly opportunistic lifestyle. This observation is in good agreement with the finding that the abundance of SAR92 was directly affected by the diatom bloom and the concurrent increase in carbohydrate concentrations [3].

This finding has to be further verified, as the CAZyme annotation pipeline is designed to identify all CAZymes in a given genome, making it prone to errors. Therefore, every single annotated CAZyme theoretically has to be manually checked for correct annotation, a process that would have been impossible in the time frame of this thesis.
High gene frequencies of GH16 and GH17 especially in the SAR92 taxobin indicates its capacity to degrade β-1,3/1,4 glucans. Laminarin is a prominent β-1,3 glucan found in brown algae like the Sargassum weed *Sargassum muticum*. This highly invasive species is speculated to have arrived in the North Sea via Japanese oysters and attached to ships [114]. It has specifically been described at the MIMAS sampling site of Helgoland since 1988 [115] and meanwhile occurs in high numbers in the North Sea.

The CARD-FISH analysis of the MIMAS study showed that the abundance of SAR92 increased fivefold after the diatom bloom. Diatoms are known to use the β-1,3 glucan chrysolaminarin as storage compounds [48]. Since the diatom bloom abruptly vanished within around 2 weeks and the SAR92 responded with a delay of around 2 weeks to this sudden decline, it seems possible that the diatoms fell prey to an early predator which allowed SAR92 to feed on the released chrysolaminarin from the disrupted diatom cells.

The complete degradation of laminarin and chrysolaminarin requires not only the degradation of β-1,3 glucans, but also the hydrolysis of β-1,6 glycosidic bonds. β-1,6 glucanases are found within GH5 and GH30. While the SAR92 taxobin does contain these two families, they are however present in lower copy numbers than GH16 and GH17. This could be explained because β-1,6-branchings occur less frequently in laminarin than the linear β-1,3 glycosidic bonds. The ratio of β-1,3 to β-1,6 linkages is 15:1 in laminarin and 11:1 in chrysolaminarin [48]. Protein trees of GH5 and GH30 did not show a close grouping of SAR92 sequences with β-1,6 glucanases (not shown in results). This could be interpreted in two ways: It is possible that SAR92 cuts out only linear β-1,3 glucan fragments and cannot process the complete (chryso)laminarin. It could however also be a matter of false or incomplete annotation within the CAZy database. For instance, *Zobellia galactanivorans* DsiJT, is able to grow on laminarin as sole carbon source [42], but does not possess an explicit β-1,6-glucanase entry in the database. It’s GH5 sequences are merely annotated as endo-beta-1,4-glucanase or simply endoglucanase.

Measurements documenting the abundance of polysaccharides in the water column after the diatom bloom would be extremely helpful to verify if these substrates were indeed still available when the abundance of SAR92 increased. Laminarin and chrysolaminarin could have also been degraded by the time SAR92 increased. The diversity in the CAZyme profile of SAR92 indicates a broad substrate spectrum, so it is possible that SAR92 utilized a different substrate than laminarin or chrysolaminarin.

A potential other substrate for SAR92 might be agar, which is an abundant component of red algal cell walls. Agar is a sulfated galactan which consists of a linear backbone of galactose residues linked by alternating β-1,4 and α-1,3 glycosidic bonds. The β-1,4
glycosidic bond can be hydrolyzed by GH16 enzymes. The α-1,3 glycosidic bond was shown to be hydrolyzed by α-1,3-(3,6-anhydro)-L-galactosidase from GH117 [42], which was present in the SAR92 taxobin (figure 10), but not in the draft genome of HTCC2207. This shows that metagenomes can cover genomic areas which have not been covered by the draft genome yet, given enough sequencing depth.

Investigation of a PUL in the genome of HTCC2207 showed a co-localization of GH16 and TonB-dependent receptors, providing further evidence to the interaction of TBDTs and CAZymes in the uptake and degradation of large polysaccharides [51, 52]. Microscopy pictures by Mariette Kassabgy showed a large fraction of SAR92 in close vicinity of phytoplankton cells or attached to macroscopic organic matter during the peak of their abundance after the diatom bloom, further indicating the clade’s affinity to oligomers and possibly high molecular weight substrate.

SusD domains, which were found in all analyzed flavobacterial species, were not found in the genome of HTCC2207, indicating that SusD might be involved in the binding of substrates not degradable by strain HTCC2207. Another possibility is that SusD is replaced by a yet unknown protein in HTCC2207. The sulfatases located in the PUL are most likely responsible for removing sulfate moieties frequently found in marine polysaccharides like agar. These moieties need to be cleaved off in order for the polysaccharide to be further processed.

Summarizing, the CAZyme analysis gave a valuable first schematic insight into the carbohydrate degradation potential of SAR92. It is however essential to keep in mind the limitations and impreciseness of CAZyme profiles and especially the potential sources of error of the CAZyme annotation pipeline used in this approach. GH families are defined by their active site motif but can have numerous substrates within one family. Therefore the definite substrate cannot be determined by CAZyme analysis alone but needs further annotation, homology prediction and molecular modeling which can provide a detailed insight into the enzyme-substrate interaction.

Nevertheless, a combination of knowledge of GH capacities, available substrates and the ecology of a specific (sampling) environment allows for a first assessment of the carbohydrate degradation potential of an organism.
4.2.2 Molecular modeling

In the GH16 phylogenetic tree, sequences from the SAR92 taxobins and from the draft genome of HTCC2207 were interspersed among laminarinas. The SAR92 sequences and the laminarinases formed a closed cluster. On one hand, this indicates that the GH16 CAZyme of SAR92 were very likely laminarinases. On the other hand, the clustering of SAR92’s GH16 CAZymes was congruent with the taxonomic classification results in the metagenomes. The proximities of the metagenomic and draft genome sequences illustrate their close taxonomic origins. In other words, the taxonomically classified SAR92 metagenomic GH16 CAZymes were likely homologous with the HTCC2207.

The modeled putative laminarinase with the best Phyre2 prediction score showed a narrow cleft topology, indicating endo activity. Endo activity has been demonstrated for Laminarinase, which showed very high selectivity for laminarin [46]. Measurement revealed that the activity was highest for laminarintetraose and slightly decreased for longer oligomers.

The active site of the putative laminarinase had a slight modification from the (so far) established pattern of laminarinases (EIDIME). The second isoleucine (I) was substituted by methionine (M), both unpolar amino acids. None of other laminarinases however come from a species even remotely related to Gammaproteobacteria, so a single amino acid mutation from one unpolar to another unpolar amino acid seems absolutely possible considering the phylogenetic distance and should also not have a considerable impact on the hydrolysis mechanism.

The putative SAR92 laminarinase was 90% identical to a laminarinase from Thermotoga maritima and received a confidence score of 100% for the prediction. During the development of the Phyre protocol, a large benchmark set of protein sequences were processed by the system and the frequency of both true- and false-positive matches was recorded. The confidence score of 100% indicates that, in the benchmark, all sequences that received this score were true homologs of the matched known protein according to the SCOP database [71]. Therefore the molecular prediction seems to be extremely accurate.

Superposition of the putative laminarinase from SAR92 and the reference laminarinase from Thermotoga maritima revealed a strikingly similar topology of large areas of the protein and especially the active site and subsites of the enzyme. The putative laminarinase showed a slightly more pronounced loop narrowing the cleft of the active site. By all other means, the active site and the subsites are however virtually identical to the confirmed laminarinase with which it was superimposed. The active site is however only responsible for the type of glycosidic bond which can be hydrolyzed. That means that different β-1,3 glycosidic bonds
can be cleaved, considering the active site pattern alone. That is why the topology of the subsites is important to investigate. Aromatic residues like the indole residue of tryptophan select for specific secondary and tertiary structures of polysaccharides. Even these indole residues superimposed well with the confirmed reference laminarinase.

Considering such a highly similar enzyme topology, it is as close as possible by in silico analysis to confirm that the putative laminarinase from SAR92 is indeed a laminarinase. In that case, SAR92 feeds on the algal polysaccharides laminarin and chrysolaminarin, occurring among the eukaryotic heterokonts Phaeophyceae (brown algae) and Bacillariophyceae (diatoms). In the context of the MIMAS study, it means that SAR92 did not necessarily thrive on substrates synthesized by the preceding bacterioplankton, but was able to utilize algal polysaccharides like laminarin, chrysolaminarin or potentially agar.

Fig. 18: Eukaryotic tree of life, depicting classes using (chryso)laminarin as storage compounds. Modified from [116]
5 Outlook

The three discovered clusters of the SAR92 clade should be confirmed by designing new FISH probes targeting those clusters exclusively. If specific probes, targeting either the pre-bloom or the bloom cluster, dominate on filters taken at those dates, the seasonality of SAR92 could be confirmed.

In silico analysis by CAZyme annotation, multiple sequence alignment with confirmed laminarinases and their superposition with the putative laminarinase showed strong indications that SAR92 possesses a laminarinase. First transcriptomic data from Illumina reads showed expression of laminarinases (Alexander Mann, personal communication) and proteomic data confirmed GH16 proteins from SAR92 in the MIMAS study [3]. It would also be interesting to see if the PUL on SAR92 strain HTCC2207 might be derived from lateral gene transfer, as no comparable loci have been found on other Gammaproteobacteria so far.

If a respective SAR92 North Sea strain could be isolated, it would likely be able to grow on (chryso-) laminarin. Similar β-1,3 linked polysaccharides of different secondary and tertiary structures should display minimal to no degradation, as shown in a laminarin degradation study by Alderkamp et al [47]. SAR92 would be the first Gammaproteobacterium growing on laminarin.

Laminarin is a major storage compound in brown algae and chrysolaminarin in diatoms. Both algae have been eyed for the production of second generation biofuels, whose energy is derived from biological carbon fixation. Algae, in contrast to land plants, do not contain lignocellulose, which, due to its heterogeneous structure, is extremely difficult to degrade. They do not require valuable land area and freshwater for growing. Furthermore, algae have been estimated to yield between 10 and 100 times more fuel per unit area than other second-generation biofuel crops [117]. Due to these advantages, government and corporate funding for the exploration of algae fuels ranges in billion dollar area.

Laminarinase is involved in the degradation of (chryso)laminarin to glucose monomers, which is necessary prior to its fermentation to ethanol, which can be used as fuel. A marine, gammaproteobacterial laminarinase could therefore be cloned into organisms like the likewise gammaproteobacterial model bacterium E.coli, which is already eyed for biofuel production [118, 119]. E.coli has recently been engineered to degrade, uptake, and metabolize alginate, another major carbohydrate constituent of brown algae [120]. The laminarinase could make the glucose monomers of the laminarin available for fermentation and subsequent biofuel production.
Finally, it should be noted that a strong focus was purposely laid on the in-depth investigation of laminarin in this thesis. The aim was the implementation of a new workflow following CAZyme annotation, aiming at a more detailed investigation of annotated enzymes through molecular modeling. The CAZyme profile of SAR92 strain HTCC2207 as well as the SAR92 taxobins showed a broad spectrum of other GH families potentially containing a plethora of enzymes involved in the degradation of other carbohydrates. These GH families should be investigated as thoroughly as the putative laminarinase in future analyses of SAR92’s carbohydrate degradation potential.
8 Acknowledgements

I would like to thank Prof. Dr. Rudolph Amann for allowing me to work in his Molecular Ecology department and providing the infrastructure for my work. His friendly nature and competent ethic undeniably carries over to his department. The seminars regularly gave me thought-provoking impulses for my own work. I also found a lot of friends in the department and would like to continue working in it.

Thank you PD Dr. Bernhard Fuchs for helping me to critically evaluate all of my results and guiding me through the phylogenetic analysis. Thank you for the helpful discussions we had in your bacterioplankton group.

Thank you Dr. Hanno Teeling for all the feedback and constructive criticism you gave me during my bioinformatic investigations, but also letting me work freely. You and the other members of the Microbial Genomics Group supported me in a research field and working methods relatively new to me.

Thank you to the Ph.D. students Sixing Huang, Johannes Werner and Alexander Mann for all the helpful advice and discussions, and introducing me to many (bio-) informatic working methods. You showed me what incredible potential bioinformatics yield, a field that often alienates traditional biologists.

Thank you my fellow MarMic classmates for the incredible time spent during the busy schedule of our master program. We all became real friends and I will never forget the time we spent together learning, but also having fun.

Last, but not least, I would like to thank my girlfriend, my family and friends for supporting and encouraging me during these sometimes stressful times and long days.
7 References


67. Chen W, Xie T, Shao Y, Chen F: **Phylogenomic relationships between amylolytic enzymes from 85 strains of fungi.** *PloS one* 2012, **7**:e49679.

68. Sigrist CJA, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I: **New and continuing developments at PROSITE.** *Nucleic acids research* 2013, **41**:D344–7.


