

Dissertation

---

Tool and Database Development for the  
Phylogenetic Classification and  
Functional Characterisation  
of Organisms

---

Dipl. Inf. Christian Quast  
September 8<sup>th</sup> 2009



**Universität Bremen**

**Fachbereich 3**



**Max Planck Institut für Marine Mikrobiologie**





**Universität Bremen**

Studiengang Informatik, FB 3  
Bibliothekstraße 1, MZH  
28359 Bremen

**Max-Planck-Institut für  
Marine Mikrobiologie**

Mikrobielle Genomforschungsgruppe  
Celsiusstrasse 1  
28359 Bremen

Doktorarbeit

zur Erlangung des akademischen Grades *Doktor der Ingenieurwissenschaften*  
an der Universität Bremen vorgelegt von Dipl. Inf. Christian Quast.

Erstgutachter: Prof. Dr. Frank Oliver Glöckner

Zweitgutachter: Prof. Dr. Otthein Herzog



## Abstract

Molecular Biology has become an integral part of every-days work in modern Biology. At the same time, sequencing technologies generate enormous amounts of genomic data in a very short time frame. Powerful bioinformatics tools are required to analyse and interpret these data. This work focuses on the development of tools for two distinct topics in the field of Bioinformatics: a pipeline to automatically build databases for the phylogenetic identification and classification of organisms, as well as a tool for the functional characterisation organisms and metagenome studies.

**Silva - Phylogenetic Classification** ARB (1) is a software workbench that is used in the ecological study of microbial communities for more than a decade. It includes tools for the phylogenetic identification of single organisms as well as tools for the design of probes to quantitatively analyse environmental samples. As such it relies on comprehensive databases of selected marker genes. In most cases, the *small subunit (SSU) ribosomal ribonucleic acid (rRNA)* is used as marker gene. Until 2004, the main ARB databases for the small subunit and for the *large subunit (LSU) rRNA* were provided by Dr. Wolfgang Ludwig (Department of Microbiology – Technische Universität München).

These manually curated databases contain high quality alignments but were limited in size and taxonomic coverage. The latest release of the SSU database (January 2004) contains approximately 40,000 sequences of all three domains of life (*Archaea*, *Bacteria*, and *Eukarya*). In 2004, this was already less than 40% of the publicly available SSU sequences contained in the databases maintained by the International Nucleotide Sequence Database Collaboration (INSDC). Due to the exponential growth of these databases, the gap between all publicly available sequences and aligned sequences can not be closed manually.

The European ribosomal RNA database (2), and the two US projects Greengenes (3), and the *Ribosomal Database Project (RDP)* (4) try to close this gap. Of these three projects, only the Greengenes project provides databases in the ARB database format. No project includes sequences from all three domains and non full-length sequences. All projects solely focus on the SSU marker gene and do not provide databases of aligned LSU genes.

In the SILVA project, a pipeline was developed to automatically create comprehensive databases including sequences from all three domains as well as non full-length sequences. This pipeline includes tasks to: extract annotated sequences, predict rRNA in otherwise not annotated environmental samples, import wrongly annotated sequences based on whitelists, check the quality of the imported sequences, align the imported sequences, and export the whole database or parts of it in various formats including the ARB database format. The SILVA project is closely tight to the ARB project at the Technische Universität München to ensure compatibility with current releases of the ARB software.

**MicHanThi - Functional Characterisation** The second part of this thesis addresses the functional characterisation of organisms and metagenome studies. Today, advancements in sequencing technology allow biologist to easily obtain the genomic sequence of a single organism, or the complete genomic content of an environmental sample. While a few years back the annotation of a single genome was the focus of several PhD students, nowadays biologists need to annotate tens of thousands of predicted genes as complement to their wetlab experiments. Tools for the automatic annotation of genes / genomes are, therefore, urgently needed.

Initial tasks in the annotation process like the prediction of potential genes (*open reading frames – ORFs*) and homology searches are automatised quite well and several specialised tools exist for each task. Stand-alone tools to infer a gene function based on the results provided by the previous tools, however, are rare. By now, most sequenc-

ing centres provide a draft annotation for the sequenced genomes. This annotation is commonly created by in-house integrated annotation systems that are not available to the public. Additionally, some institutes provide web-based solution. Examples are the *Rapid Annotation using Subsystem Technology (RAST)* (5) hosted by the Mathematics and Computer Science Division at the Argonne National Laboratory and the *Integrated Microbial Genomes (IMG)* genome browsing and annotation system developed by the Department of Energy (DOE) Joint Genome Institute (JGI) (6).

Most of these systems provide only limited or no control over the annotation pipeline and do not give reliability scores for the predicted annotations. This hampers biologists during the post processing of this data – whether or not to trust the predicted functions. Another important aspect is that these systems cannot be installed locally, further limiting their use in academic and particularly industrial projects.

MicHanThi focuses on the prediction of gene functions based on the results of tools such as BLAST (7) and InterProScan (8). Rather than running these tools itself, MicHanThi relies on the results stored in the GenDB annotation system (9). In this thesis, the prototype developed in (10) was enhanced to include InterPro (11) domain predictions as well as to utilise the relationship among InterPro entries.

# Contents

<b>I</b>	<b>Introduction</b>	<b>1</b>
<b>1</b>	<b>Bioinformatics</b>	<b>3</b>
1.1	Generation and Analysis of Molecular Data . . . . .	6
1.1.1	Sequencing . . . . .	6
1.1.2	Assembly and Binning . . . . .	8
1.1.3	ORF prediction . . . . .	8
1.1.4	Functional Annotation . . . . .	9
1.1.5	Phylogenetic Classification . . . . .	9
1.2	Methods . . . . .	10
1.2.1	Sequence Alignment . . . . .	10
1.2.2	Pattern / Profile (Motif) Searches . . . . .	12
1.3	Databases . . . . .	13
1.3.1	Relational Databases and SQL . . . . .	13
1.3.2	Sequence Databases . . . . .	15
1.3.3	Pattern / Profile Databases . . . . .	17
1.3.4	rRNA Databases . . . . .	19
1.4	Sequencing Artifacts . . . . .	19
1.4.1	Vectors and Vector Contamination . . . . .	19
1.4.2	Chimeras and Chimera Detection . . . . .	20
<b>2</b>	<b>Research Objectives</b>	<b>21</b>
2.1	Phylogenetic Classification . . . . .	21
2.2	Functional Characterisation . . . . .	22
<b>3</b>	<b>Summary</b>	<b>23</b>
3.1	SILVA . . . . .	24
3.1.1	Tasks . . . . .	25
3.1.2	Web Presence . . . . .	29
3.1.3	Design and Implementation . . . . .	31
3.2	MicHanThi . . . . .	34
3.2.1	Process Flow . . . . .	35
3.2.2	Results . . . . .	38
3.2.3	MicHanThi Accuracy / Human Inaccuracy . . . . .	39
<b>4</b>	<b>Of Avalanches and Tsunamis</b>	<b>41</b>
4.1	Homology Searches . . . . .	42
4.2	ORF Prediction . . . . .	43
4.3	Representative Sets . . . . .	44

---

4.4	Conclusions . . . . .	46
<b>5</b>	<b>Acknowledgments</b>	<b>49</b>
<b>II</b>	<b>Publications</b>	<b>51</b>
<b>6</b>	<b>Silva Paper</b>	<b>55</b>
<b>7</b>	<b>MicHanThi Manuscript</b>	<b>71</b>
<b>8</b>	<b><i>Gramella forsetii</i> KT0803 Paper</b>	<b>91</b>
<b>9</b>	<b><i>Congregibacter litoralis</i> KT71 Paper</b>	<b>105</b>
<b>10</b>	<b>Pirellula Paper</b>	<b>119</b>
<b>11</b>	<b>Megx.net Paper</b>	<b>141</b>
<b>III</b>	<b>Appendix</b>	<b>149</b>
<b>A</b>	<b>Tools, Libraries &amp; Databases</b>	<b>151</b>
<b>B</b>	<b>MicHanThi Rule Base &amp; SILVA Meta Data</b>	<b>153</b>
<b>C</b>	<b>MicHanThi Design</b>	<b>157</b>



# List of Abbreviations

<i>C. litoralis</i> ...	<i>Congregibacter litoralis</i> KT71
<i>E. coli</i> .....	<i>Escherichia coli</i>
<i>G. forsetii</i> ...	<i>Gramella forsetii</i> KT0803
<i>H. influenzae</i>	<i>Haemophilus influenzae</i>
<i>O. algarvensis</i>	<i>Olavius algarvensis</i>
<i>R. baltica</i> ....	<i>Rhodopirellula baltica</i> SH1 <sup>T</sup>
ABI .....	Applied Biosystems
API .....	Application Programming Interface
BASH .....	Bourne-again shell
BLAST .....	Basic Local Alignment Search Tool
BMBF .....	Bundesministerium für Bildung und Forschung – Federal Ministry of Education and Research
CDS .....	Coding Sequence
DBMS .....	Database Management System
DDBJ .....	DNA Data Bank of Japan
DNA .....	Deoxyribonucleic Acid
DOE .....	Department of Energy
DSMZ .....	Deutsche Sammlung für Mikroorganismen und Zellkulturen – German Collection of Microorganisms and Cell Cultures
E-value .....	Expect Value or Expectation Value
EC number ..	Enzyme Commission number or Enzyme Classification number
EMBL .....	European Molecular Biology Laboratory
GG .....	Greengenes
GI number ...	The Unique Identifier in the GenBank Database
GiB .....	Gibibyte – 2 <sup>30</sup> Bytes
GO number ..	Gene Ontology number
GSA .....	Global pairwise Sequence Alignment
HTU .....	Hypothetical Taxonomical Unit
IMG .....	Integrated Microbial Genomes
INSDC .....	International Nucleotide Sequence Database Collaboration
JGI .....	Joint Genome Institute
KB .....	Kilo Bases – One Thousand Bases
KiB .....	Kibibyte – 2 <sup>10</sup> Bytes
LSA .....	Local pairwise Sequence Alignment
LSU .....	Large Subunit
MB .....	Mega Bases – One Million Bases
MiB .....	Mebibyte – 2 <sup>20</sup> Bytes
MIMAS .....	Microbial Interactions in Marine Systems

---

MPI	.....	Massive Parallel Instruction
mRNA	.....	Messenger RNA
MSA	.....	Multiple Sequence Alignment
NCBI	.....	National Center for Biotechnology Information
NIH	.....	National Institute of Health
ORF	.....	Open Reading Frame
OTU	.....	Observed Taxonomical Unit
PCR	.....	Polymerase Chain Reaction
PDB	.....	Protein Data Bank
PIR	.....	Protein Identification Resource
PRF	.....	Protein Research Foundation
PT	.....	Positional Tree
RAST	.....	Rapid Annotation using Subsystem Technology
RCSB	.....	Research Collaboratory for Structural Bioinformatics
RDBMS	.....	Relational Database Management System
RDP	.....	Ribosomal Database Project
RefSeq	.....	Reference Sequence
RNA	.....	Ribonucleic Acid
rRNA	.....	Ribosomal Ribonucleic Acid
SINA	.....	SILVA INcremental Aligner
SQL	.....	Structured Query Language
SSU	.....	Small Subunit
STL	.....	Standard Template Library
TiB	.....	Tebibyte – $2^{40}$ Bytes
TREMBL	.....	Translated EMBL
TUM	.....	Technische Universität München
UniProt	.....	Universal Protein Resource
UniProtKB	..	Universal Protein Resource Knowledge Base

# List of Figures

1.1	The Full-Cycle rRNA Approach . . . . .	5
1.2	Sequence alignment of two random unaligned sequences. . . . .	11
1.3	Extract from a multiple sequence alignment of five ORFs coding for ‘serine protease do-like precursor’ (degP). . . . .	12
1.4	Prokaryotic membrane lipoprotein lipid attachment site. . . . .	13
1.5	An exemplary profile. . . . .	14
1.6	A simple database schema that can be used to model the data in genome annotation projects. . . . .	15
1.7	Growth of Sequence and 3D Structure Databases. . . . .	16
1.8	Example of the description of an entry in the NCBI nr database ( <i>gi 16121437 ref NP_404750.1 </i> ). . . . .	17
3.1	Workflow and interactions in the SILVA pipeline. . . . .	26
3.2	The SILVA web presence at <a href="http://www.arb-silva.de">http://www.arb-silva.de</a> . . . . .	30
3.3	SILVA database design. . . . .	32
3.4	The MicHanThi annotation process . . . . .	35
3.5	Fuzzy Logic membership functions . . . . .	37
4.1	Number of transistors used in Intel Desktop CPUs and the growth of SSU rRNA sequence databases. . . . .	42
4.2	Two ORF predictions of <i>G. forsetii</i> and contig contig00408 of the Logatchev metagenome study (viewed in GenDB) . . . . .	44
6.1	Sequence length distribution of rRNA genes in the SILVA 91 SSU database. . . . .	67
6.2	Sequence length distribution in the SILVA 91 LSU database. . . . .	67
7.1	The Annotation Process. . . . .	76
7.2	Definition of the member functions for the linguistic variables of BLAST and InterProScan. . . . .	78
7.3	Splitting observation descriptions into atoms . . . . .	80
7.4	Generation of groups. Each group contains a list of atoms and all observations containing each atom in its description. . . . .	81
8.1	Comparison of hydrolytic capabilities and adhesion potential. . . . .	95
8.2	Comparison of gene family profiles. . . . .	96
8.3	Comparison of abundance and types of proteins potentially mediating surface adhesion. . . . .	101

---

9.1	Phylogenetic affiliation of KT71. . . . .	108
9.2	Comparison of PS operons. . . . .	109
9.3	Pigment analysis. . . . .	110
10.1	Number of regulated genes per stress experiment. . . . .	123
10.2	Number of regulated genes with an assigned COG-category. . . . .	124
10.3	Venn diagrams of specific and common stress response. . . . .	125
11.1	Fast access to the annotation highlights of marine microorganisms. . . . .	145
11.2	The Genomes Mapservers. . . . .	147
C.1	MicHanThi modules overview. . . . .	157
C.2	Module IO overview. . . . .	158
C.3	Module DATA overview. . . . .	158
C.4	Module TOOL overview. . . . .	159
C.5	Module ANNOTATOR overview. . . . .	159

# List of Tables

1.1	Evolution of next generation sequencers. . . . .	7
6.1	Description of database fields in ARB files exported from SILVA for ARB specific fields and entries. . . . .	59
6.2	Description of database fields in ARB files exported from SILVA for Fields and entries imported from EMBL. . . . .	61
6.3	Description of database fields in ARB files exported from SILVA for SILVA specific fields and entries. . . . .	64
6.4	Sequence retrieval and processing for SILVA 91 . . . . .	66
7.1	Overall statistics of the comparison of annotation created by human annotators and annotations created by MicHanThi. . . . .	83
7.2	Detailed comparison of annotations for ORFs without a functional assignment. . . . .	84
7.3	Overall statistics of the comparison of the revised human created annotations and annotations created by MicHanThi. . . . .	85
7.4	Details of the mismatches in classes hypothetical, conserved, and domain. . . . .	86
7.5	Sementatically equivalent annotation created by the human annotator MicHanThi. . . . .	89
8.1	General features of the 'Gramella forsetii' KT0803 genome. . . . .	94
10.1	Shared stress response to heat, cold and high salinity: Results for induced genes are shown. . . . .	134
10.2	Shared stress response to heat, cold and high salinity: Results for repressed genes shown. . . . .	136
10.3	Differentially expressed sulfatase genes of R. baltica are shown. . . . .	139
A.1	Resources used in this thesis. . . . .	151
B.1	Rule base used to evaluate the reliability of BLAST observations. . . . .	153
B.2	Meta data exported into the ARB database files and their sources. . . . .	154



## Part I

# Introduction





# Chapter 1

## Bioinformatics

In the late 19th century, Robert Koch was the first to apply pure culture techniques to study microorganisms. Since then, biologists are studying the metabolic capabilities, resistance, and pathogenesis of microorganisms on isolates.

A new era of Biology was entered when Francis Crick and James Watson discovered the *Deoxyribonucleic acid (DNA)* (12). Since the first genome was completely sequenced in 1977 by Frederick Sanger (13) (*Bacteriophage  $\phi$ -X174*), numerous sequencing projects were successfully accomplished, including projects such as the first microorganism (*Haemophilus influenzae*) (14) and the human genome project (15). Today, the complete or nearly complete genomic sequence of more than 1,000 organisms is known<sup>1</sup>. Additionally, the genomes of more than 3,600 organisms are currently sequenced. The extraction of DNA, genome mapping, data storage, and computer aided analysis of the data became known as *Genomics*. Today, the sequence data of all published genes and genomes is stored in public databases hosted by the International Nucleotide Sequence Database Collaboration (INSDC). This consortium is a collaboration between the DNA Data Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL) and National Institute of Health (NIH).

In 1990 Torsvik (16) was one of the first to introduce culture-independent methods to investigate the diversity of microorganisms. Her study revealed a high diversity on the DNA level although previous phenotype based classification suggested otherwise. Today, it is believed that less than one percent of the organisms can be cultivated using common techniques.

Large sequencing capacities, advancements in sequencing technologies, and the possibility to study organisms independent of pure cultures have lead to a paradigm shift in biology. Instead of concentrating on the study of single, cultivated organisms using their closed genomes, biologists now focus on the study of genomic fragments directly extracted from environmental samples (*metagenomics*) (17). In 1996, Stein was among the first to publish a metagenomic library (18).

In *molecular microbial ecology* these culture-independent techniques are routinely used to answer the questions which organisms are in the environment, how many of these organisms are there (community structure), and in which processes are these organisms involved (what are they doing / what is their function). Data processing, especially considering the enormous sequencing capacities and there-

---

<sup>1</sup>Genomes OnLine Database (GOLD), June 2009; <http://www.genomesonline.de>

fore available sequence information, is still in an early stage. Databases and tools to handle these data masses are urgently needed and need to be developed. These tools will then help biologist answer the postulated questions.

**Phylogenetic Classification** Biologists typically classify animals based on their phenotype. Animals are believed to be related if they share certain traits (e.g. number of legs, colour / pattern of fur, size). They are more closely related if they share more traits (forming kingdoms, classes, and families). This type of classification is normally applied to mammals, birds, and other animals with distinctive phenotypes. For microorganisms (e.g. *Bacteria*) this does not work because only few observable phenotypes exist (16).

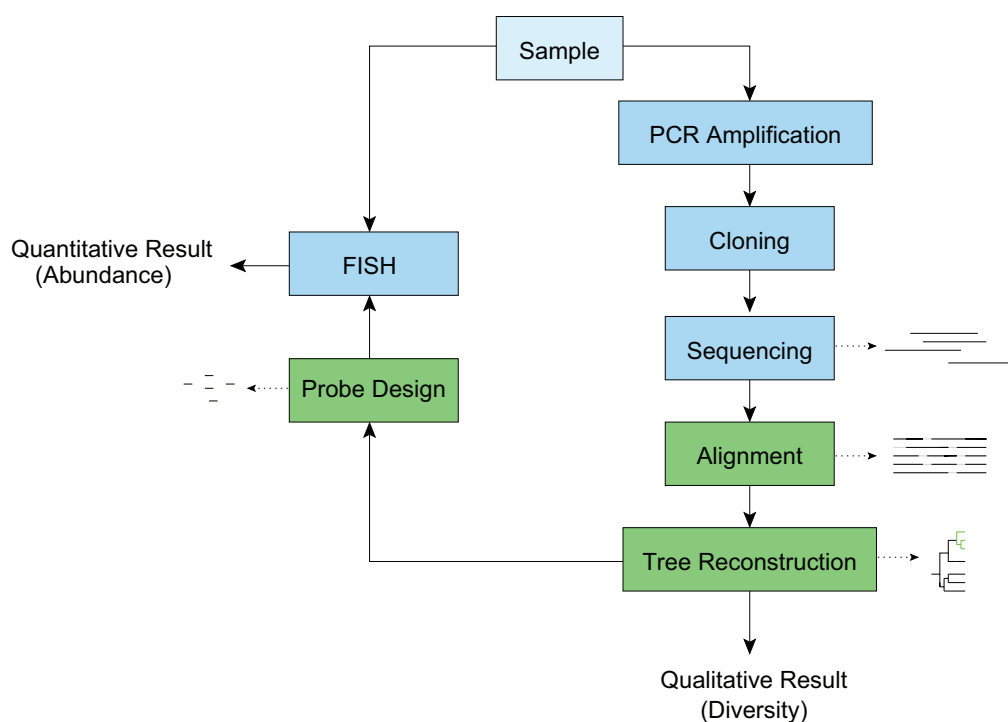
Today, *Bacteria* and other microorganisms are usually identified by comparing certain regions of their DNA (*marker genes*). These marker genes evolved over time and still share the same function (*orthologous*). The degree of relationship of two organisms is then defined by the evolutionary distance of these marker genes. Most studies to classify organisms are based on the *small subunit ribosomal RNA (SSU rRNA)* marker gene.

After a sample is taken, a selected marker gene is specifically amplified by *polymerase chain reaction (PCR)* from the genomes of all organisms in the sample. Clone libraries are created to separate all variants of the marker gene within the sample, which are then sequenced. Based on phylogenetic reconstruction the sequenced DNA is compared to known copies of the same gene in the sequence databases. Subsequently, molecular probes can be designed for selected phylogenetic groups in the tree.

A probe is a sequence signature representing a sub region on the marker gene which uniquely identifies a certain species or group of organisms. Additionally, fluorescence markers are appended to the probe for subsequent wetlab experiments. These probes are then used *in situ* to specifically stain the organisms in a sample. To quantify the different types of microorganisms and to quantify their numbers fluorescence microscopy is used. This process of organism identification and quantification is known as the *full-cycle rRNA approach* (19) (Figure 1.1).

Comprehensive, high quality databases of aligned marker genes are essential to assure both the sensitivity and the specificity of probes. If a group of organisms in the tree does not suitably represent all members of that group, false sensitivity might be assumed. This may lead to an under representation of members of this group in a sample. While missing members of a group may influence the sensitivity of a probe, a missing group of organisms may reduce its specificity. In a molecular study this may cause over representation of the group of organisms represented by the probe.

Until 2004, databases of manually aligned of the *large subunit (LSU)* and *small subunit* rRNA genes were provided, amongst others, by Dr. Wolfgang Ludwig (Department of Microbiology – Technische Universität München; TUM). Due to advancements in sequencing technology and the reduction of sequencing cost the number of available rRNA genes increases exponentially. It has reached the number of one million available sequences in the LSU and SSU databases. Providing manually curated databases is, in consideration of this development, not feasible. Systems to automatically provide comprehensive, quality controlled databases of aligner rRNA sequences are urgently needed.



**Figure 1.1:** *The Full-Cycle rRNA Approach modified after (20). Tasks applied to identify and quantify organisms and organism groups in biological samples. Blue in situ tasks, green in silico tasks.*

**Functional Characterisation** Another important aspect in Genomics and Metagenomics, besides the identification and quantification of microorganisms, is their functional characterisation. This is necessary to answer ecological questions concerning what single microbes are doing, how they are doing it, how they interact with their environment and how they interact with other organisms within microbial communities.

After the genomic information has been extracted from a sample and it has been assembled, open reading frames (ORFs) can be predicted. These ORFs are potential genes and their sequence is compared to publicly available gene databases. Once all ORFs have been functionally characterised, the metabolic capabilities of an organism can be reconstructed and a life style can finally be predict.

In 2003, Glöckner et. al published the annotation of the marine organism *Rhodospirillum rubrum* SH1<sup>T</sup> (21). The manual annotation of the approximately 7,300 submitted genes and the metabolic characterisation of this organism took more than three years. Today, the screening of a single environmental sample and especially metagenomic studies e.g. (22) reveal a multitude of ORFs. While a few years back the annotation of an organism was the joined work of a group of researchers, nowadays, the functional characterisation using Bioinformatics methods is considered to be a complement to wet lab studies. Automatic tools to support the biologist in the study of these data masses are urgently needed.

## 1.1 Generation and Analysis of Molecular Data

The main aspect of this work is the development of tools to support the biologist in the molecular study of biological samples. One tool of the developed tools supports the biologist in the functional study of organisms by proposing functions for predicted genes. The second set of tools was developed to automatically create databases for LSU and SSU rRNA marker genes that are used in phylogenetic studies.

The following sections describe the tasks that need to be conducted in the functional and phylogenetic study of organisms and complete biological samples.

### 1.1.1 Sequencing

Sequencing is the initial task in the genomic study of organisms and environmental samples. Its purpose is to extract the DNA contained in a biological sample and to make the DNA available for the analysis by the computer.

Since the first sequencing projects in the late 1970's the most widely used method for sequencing was Sanger sequencing. Later, this method was complemented by the *Shotgun approach* to make it applicable to the sequencing of complete genomes. The underlying method of *dideoxy chain termination* stayed mostly unchanged over the years. In the early 1990's, the time needed for sequencing could be reduced drastically by the introduction of new sequencing strategies and the introduction of capillary sequencers, but the cost for sequencing remained high. These new systems, however, were still based on the methods initially developed in the 1970's. More than two decades passed, until the turn of the millennium, before fundamental changes were made to the methods underlying sequencing.

In 2001, Ronaghi published an article on advancements in sequencing technology (23). He describes a newly developed method called *pyrosequencing* that rigorously breaks with the older concepts used by Sanger sequencing. Instead of sequencing by electrophoresis this method follows the sequencing by synthesis approach. 454 Life Sciences, which is now owned by La Roche Ltd, licensed this technique and adapted it for large-scale sequencing projects (24). In less than 5 years, 454 Life Sciences developed three generations of sequencers based on pyrosequencing. With each generation, the sequencing throughput and the average read length could be increased while the cost for sequencing could be reduced at the same time.

Besides 454-pyrosequencing, two more sequencing robots based on sequencing by synthesis are currently in the market, the *Illumina / Solexa Genome Analyzer II (GA II)* (25), and the *SOLID 2* system developed by Applied Biosystems (26). All three so called *next generation* sequencing methods increased the throughput thousandfold compared to the older Sanger sequencing. The relatively long average read length of Sanger sequencing of up to 800 bases could, however, not be retained. Currently, 454 Titanium (Ti) sequencers reach an average read length of approximately 350 bases, leaving the competing next generation sequencing methods far behind, GA II 75 bases and SOLID 2 35 bases.

A summary of the capabilities of the next generation sequencing techniques can be found in Table 1.1.

## Next Generation Sequencing Statistics

Vendor:	Roche			Illumina			ABI	
Technology:	454			Solexa			SOLiD	
Platform:	GS 20	FLX	Ti	GA	GA II		1	2
Reads: (M)	0.5	0.5	1	28	100		40	115
<b>Fragment</b>								
Read length:	100	200	350	35	50	75	25	35
Run time: (d)	0.25	0.3	0.4	3	3	4.5	6	5
Yield: (GB)	0.05	0.1	0.4	1	5	7.5	1	4
Rate: (GB/d)	0.2	0.33	1	0.33	1.67	1.67	0.34	1.6
Images: (TB)	0.01	0.01	0.03	0.5	1.1	1.7	1.8	2.5
PA Disk: (GiB)	3	3	15	175	300	350	300	750
PA CPU: (hr)	10	140	220	100	70	100	NA	NA
SRA: (GiB)	0.5	1	4	30	50	75	100	140
<b>Paired-end</b>								
Read length:		200		2×35	2×50	2×75	2×25	2×35
Insert: (KB)		3.5		0.2	0.2	0.2	3	3
Run time: (d)		0.3		6	10	15	12	10
Yield: (GB)		0.1		2	9	12	2	8
Rate: (GB/d)		0.33		0.33	1.67	1.67	0.34	1.6
Images: (TiB)		0.01		1	2.2	3.4	3.6	5
PA Disk: (GiB)		3		350	500	600	600	1500
PA CPU: (hr)		140		160	120	170	NA	NA
SRA: (GiB)		1		60	100	150	200	280

**Table 1.1:** *Evolution of next generation sequencers.*

*ABI Applied Biosystems; PA is primary analysis (includes image feature extraction and base calling); PA CPU is calculated as the wall clock multiplied by the number of CPU cores; ABI SOLiD data, except rate, are representative of a single slide; ABI SOLiD primary analysis is done on the instrument cluster; SRA is the size of the files (SFF or SRF) that are submitted to the NCBI Short Read Archive;*

*Source: <http://www.politigenomics.com/next-generation-sequencing-informatics>*

### 1.1.2 Assembly and Binning

Currently, sequencing technologies applied for genome sequencing cannot sequence complete genomes as one read. Instead they produce thousands or even millions of short reads which need to be arranged in the correct order (*assembly*). This is done by arranging the reads according to overlapping parts. Ideally, all reads can be arranged and the genome is closed.

Sanger sequencing allows the assembly of complete genomes. These closed genomes commonly are of high quality. Of the next generation sequencing methods only 454-pyrosequencing can be used for the sequencing of genomes. Due to its shorter average read length, the assembly of genomes is more difficult leaving thousands of fragments which cannot be assembled (*contigs*). The other next generation techniques are normally used in genome re-sequencing and mapping, as well as *single nucleotide polymorphisms* (*SNP*) detection.

Currently, most projects related to the study of environmental samples favour 454-pyrosequencing as sequencing method. In an environmental sample, the DNA of an unidentified number of organisms is contained. Considering the problem of short read length and the difficulties to assemble these reads leaves the majority of reads unassembled or assembled to contigs of a few thousand KB. Contigs longer than one hundred KB are the exception.

To get feeling of which organisms were sequenced, the intrinsic signal of the DNA is analysed and the reads are grouped in artificial organisms bins (*binning*). These organism bins are then studied as closed genomes would be. One of the first studies to apply this approach was the study of the organism *Olavius algarvensis* (27).

### 1.1.3 ORF prediction

Once the DNA is extracted and the reads are assembled, tools are applied to predict possible protein-encoding sequences (*open reading frames – ORFs*). This is the part of a gene that is transcribed to mRNA and later translated into a protein (28).

The position at which the transcription is stopped (stop codon) is unambiguously defined by one of the triplets ‘TAA’, ‘TAG’, or ‘TGA’. However, the triplet coding for the start of the ORF (start codon) is ambiguous. In most cases, the start codon is the triplet ‘ATG’ but it can be other triplets. Furthermore, the transcription process is not always started if a start codon is encountered since ‘ATG’ also codes for the amino acid *methionine*. Hence, a lot of effort is exerted to correctly predict the start of an ORF.

Since the prediction of the start position is ambiguous, tools either predict too many ORFs or only the most likely ORFs (over prediction vs. quality of the predicted ORFs). In the annotation of the organism *Rhodospirellula baltica* SH1<sup>T</sup> more than thirteen thousand ORFs had been predicted. Of these ORFs, approximately 7,300 were finally submitted to EMBL. For more than 50% of the submitted ORFs no homologue sequences could be found in public sequence databases, at that time.

To increase the quality of the predicted ORFs and to reduce the manual work load Jost Waldmann and Dr. Hanno Teeling (Microbial Genomics Group – Max Planck Institute Bremen) developed a meta ORF finder (*MORFind*). It combines

the results of different ORF prediction tools and creates a non-redundant list of ORFs. Overlapping ORFs are considered to be contradictions in the ORF prediction and a sophisticated reasoning process is applied to solve discrepancies.

#### 1.1.4 Functional Annotation

Gene annotation is the process to associate certain information with the predicted ORFs describing their function. Among this information is: the function of the protein, a short “unique” name describing the function (*gene name*), and the classification of the ORF. The classification of an ORF can be done using different schemes. The more popular schemes are EC numbers, which classify the ORF corresponding to its metabolic pathway (29), as well as GO Numbers which classify the ORF according to its molecular function, cellular component, and biological process (30).

After the ORF prediction, the possible genes are not annotated. To derive a function for a particular ORF, its sequence is compared to already annotated genes in sequence databases (Sequence Alignment 1.2.1). Additionally, tools can be used to assign an ORF to a certain protein family by matching its sequence to patterns or profiles describing one of the currently known protein families (Pattern / Profile Searches 1.2.2).

Two homology based methods are used to automatically transfer functional annotations from previously characterised genes to novel sequences: horizontal and vertical annotation.

Horizontal annotation focuses on the annotation of single ORFs, mostly neglecting neighbouring ORFs. Two methods are commonly used to derive evidences for the functional annotation: pairwise sequence alignment (PSA), and profile hidden Markov models (HMM). PSA creates an alignment of a novel sequence and a known sequence contained in a database. A tool widely used for this type of searches is BLAST (31). The HMM based approach creates a profile from a multiple sequence alignment (MSA) which represents a group of closely related genes with the same function. It then uses this profile as a scoring schema to create a pairwise alignment (32). Two commonly used systems using the horizontal annotation approach are AutoFACT (33) and BASys (34).

Vertical annotation uses the order of genes to predict a function for a set of newly sequenced genes retaining the same order. Subsystems can further be used to enhance this annotation method. Subsystems are commonly based on metabolic pathways but may resemble any expert defined group of genes. Systems using this annotation approach, commonly fall back to horizontal annotation if an ORF cannot be annotated otherwise. The first annotation systems to facilitate this annotation approach were Ergo (35) and the SEED (36). Today, the most commonly used system is the RAST web service (5).

#### 1.1.5 Phylogenetic Classification

The phylogenetic classification of organisms typically involves building a “tree of life”. This tree represents the evolutionary relationships among organisms or other entities, such as a set of functionally related genes, that are believed to have a common ancestor. In a phylogenetic tree, each leaf represents an entity whose DNA could be obtained through sequencing, *operational taxonomic units*

(*OTUs*). Each internal node forms the most common ancestor of the nodes directly beneath it. Internal nodes are often called *hypothetical taxonomical units* (*HTUs*) as they cannot be directly observed. In some trees, depending on the algorithm used to build the tree, the length of a branch denotes to the evolutionary distance, e.g. the number of character changes, between the descendants of a node. Trees showing the evolution of the same group of organisms may differ if unrelated types of input data are used (morphological data vs. genomic data).

Phylogenetic trees may, depending on the algorithm, be created based on existing multiple sequence alignments or evolutionary models. Building a phylogenetic tree is considered NP-hard. Two types of trees can generally be distinguished: rooted trees, and unrooted trees. A rooted tree is a tree with a single root node. The tree is directed with respect to time and the root node corresponds to the most recent common ancestor of the entities at the leaves. The unrooted tree does not have a unique root node. It is not directed and without making assumptions about common ancestry shows the relationship between the entities at the leaves.

A phylogenetic tree may always only represent a hypothesis about the evolutionary ancestry of the studied entities because the evolutionary process cannot be directly observed.

## 1.2 Methods

A central aspect of Bioinformatics is the alignment of two or more sequences. The alignment is used to estimate the evolutionary distance of the sequences in the alignment. Two types of alignment are commonly distinguished: the pairwise alignment of two sequences, and the multiple sequence alignment. The pairwise alignment is used to compare a predicted ORF to known proteins in a database. Its goal is to derive a function for the newly predicted gene. Multiple sequence alignments are used to build patterns and profiles of groups of closely related genes. These patterns and profiles are then again used to functionally describe an unknown ORF. Multiple sequence alignments are also used in the phylogenetic study of organisms and the evolution of single protein family.

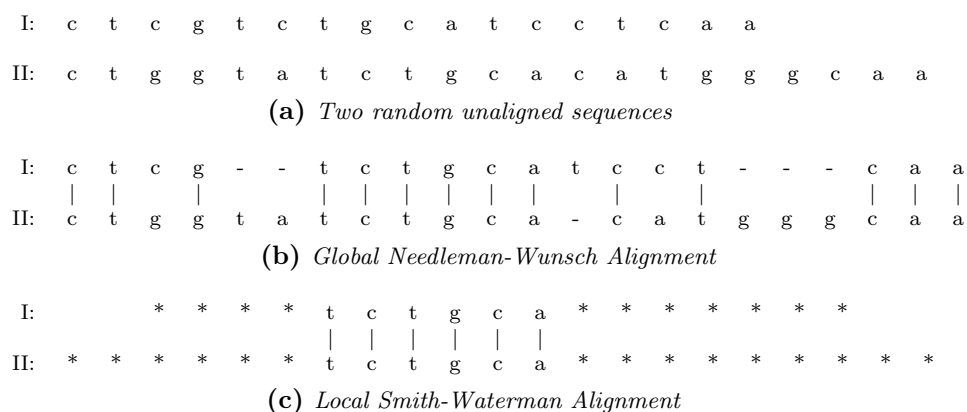
The following sections give a short overview of the concepts and of the tools commonly used to align sequences.

### 1.2.1 Sequence Alignment

Sequence alignment is a scheme of writing two or more strings on top of each other where the characters in one position are deemed to have a common evolutionary origin (*positional homology*). The algorithms developed to compare two strings are mostly based on the concepts of dynamic programming developed by Richard Bellman in the 1950s (37). These concepts refer to a multi-stage decision making process that yields optimal results and were initially not related to string analysis or the comparison of biological sequences.

In Bioinformatics, this approach is used to compare two or more DNA or protein sequences, highlighting their similarities in order to identify a common function or a common evolutionary origin. The sequences are arranged so that, when ever possible, identical bases are placed on top of each other in the align-





**Figure 1.2:** Sequence alignment of two random unaligned sequences.

ment. If necessary, gaps (usually denoted by dashes ‘-’) are introduced into the alignment. Gaps are considered to be deletions or insertions in the evolutionary process of a gene, whereas mismatches correspond to mutations. Broadly, two types of pairwise sequence alignments can be distinguished, *global sequence alignment (GSA)* and *local sequence alignments (LSA)*. The global alignment and the local alignment of two random sequences (Fig. 1.2a) are shown in Figures 1.2b and Figure 1.2c.

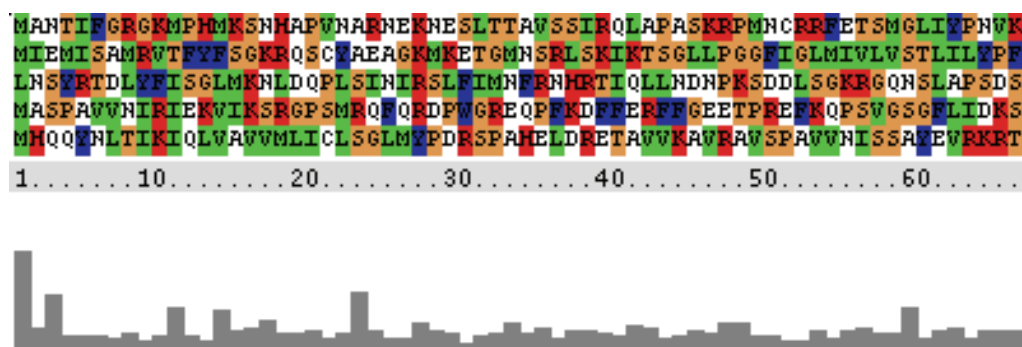
A global sequence alignment of two sequences is an alignment that spans along their entire length. Gaps are introduced as necessary to make up for the differences in length of the two sequences. Global sequence alignment is most useful for aligning and finding closely related sequences. The Needleman-Wunsch algorithm (38) was the first algorithm to apply the concepts provided by dynamic programming for the alignment of biological sequences.

An algorithm for the local alignment of two sequences was developed by T. F. Smith and M. S. Waterman in the early 1980s (39). It can be used to find closely matching regions of much longer sequences. The position of the matching regions within their parent sequences is irrelevant. This makes local sequences alignment robust against evolutionary events such as *domain shuffling*.

Pairwise sequence alignment is used to derive a function for an unidentified query sequence. A sequence is iteratively aligned against all sequences in a database containing previously annotated genes. Genes in this database may or may not be functionally described. All matches that meet a certain threshold are returned. Users may then use these results to derive a function for the unknown sequence. The more matches an alignment of two sequences shows, the better the alignment and a function may be predicted more reliably. BLAST (31) is the most commonly used for the alignment of two sequences.

An extension of the pairwise sequence alignment is the *multiple sequence alignment (MSA)* used to align more than two sequences. Multiple sequence alignment is computationally difficult and is classified as an NP-Hard problem. The most known algorithm to create multiple sequence alignments is CLUSTAL (40). Further commonly used programs include MAFFT (41) and MUSCLE (42). Figure 1.3 shows the multiple sequence alignment of five sequences using the CLUSTAL algorithm.

Multiple sequence alignments are used in the functional characterisation of



**Figure 1.3:** Extract from a multiple sequence alignment of five ORFs coding for ‘serine protease do-like precursor’ (*degP*). The CLUSTAL algorithm (40) was used to create the alignment.

organisms as well as in their phylogenetic classification and quantification. In the functional characterisation, a pattern or profile is created from the MSA of functionally related proteins. A novel sequence is then compared to the pattern or profile of this group instead of directly aligning it against all sequences. Pre-calculated MSAs are also used by some algorithms to construct phylogenetic trees.

**BLAST** The *Basic Local Alignment Search Tool* (BLAST) (31) algorithm is the most widely used algorithm for the local alignment of two sequences. Unlike the Smith-Waterman algorithm, it returns a number of statistically significant alignments rather than just the “best” one. Another difference between the two algorithms is that the Smith-Waterman is guaranteed to find the optimal local alignment between two sequences while BLAST uses a heuristic to reduce the search space. Using a heuristic increases the search speed at the cost of sensitivity. This means that an optimal alignment between two sequence may not be found.

A measure for the statistical importance of the alignment returned by the BLAST algorithm is the *Expectation value* (or *Expect value*) short E-value (43). This is the number of alignments expected by chance  $E$  during a sequence database search of search space  $m \times n$ , where  $m$  denotes the length of the query sequence and  $n$  is the size of the database in characters (the length of the concatenation of all of sequences within the database).

### 1.2.2 Pattern / Profile (Motif) Searches

Pattern or profile searches are also applied to functionally characterise proteins as is pairwise sequence alignment. Unlike pairwise sequence alignment, pattern and profile searches do not compare two sequences directly. Instead they compare a query sequence to a pattern or profile describing a domain or family of proteins. Patterns and profiles describe conserved regions in a group of genes. These conserved regions can be found by creating a multiple sequence alignment of all members of the protein family or all proteins carrying the same domain. When using patterns or profiles to characterise an unknown sequence, the conserved

regions are searched for in the query sequence. This approach seems worthwhile because different domains of a protein are subject to different selective pressures (32). This means that some parts of a protein are more conserved among a group of proteins than others.

Patterns are regular expressions describing each position of the MSA that is relevant to identify a protein family. Each position of the pattern represents one or more characters of the alphabet that are observed at the position of the MSA. Only these characters are allowed to occur at that position in a query sequence. If a character is found in the query sequence, that is not represented by the pattern for that particular position, then the query sequence is called a mismatch. This problem can be solved by allowing a number of mismatches within the pattern. An example of a pattern is shown in Figure 1.4.

{DERK}(6) - [LIVMFWSTAG](2) - [LIVMFYSTAGCQ] - [AGS] - C

**Figure 1.4:** *An exemplary pattern of the Prokaryotic membrane lipoprotein lipid attachment site: C is the lipid attachment site. Additional rules: (1) The sequence must start with Met. (2) The cysteine must be between positions 15 and 35 of the sequence in consideration. (3) There must be at least one Lys or one Arg in the first seven positions of the sequence. Source: <http://www.expasy.org/cgi-bin/nicedoc.pl?PDOC0013>.*

Profiles like Patterns describe conserved regions of a MSA. Unlike patterns, profiles specify for each position within the conserved region the probability for each character of the input alphabet by which it may occur at that particular position. Hence, profiles implicitly allow mismatches at any given position of the profile because it is “just” more likely for some characters to occur at a certain position of the alignment. Characters which do not occur in the MSA are assigned a probability close to zero. This means that it is very unlikely that one of those characters will occur. Algorithms implementing *hidden Markov Models* are most commonly implemented to create such profiles. An exemplary profile is depicted in Figure 1.5

## 1.3 Databases

Databases of various types of information play a central role in Bioinformatics. In functional Genomics, each newly predicted gene is compared to databases of already known genes, as well as to databases containing patterns and profiles describing functionally related proteins. Databases of the rRNA marker gene are used to reconstruct the evolutionary relatedness between organisms. Also, these databases are used to design probes which are used to identify organisms in biological samples. Most databases in Bioinformatics use relational databases.

### 1.3.1 Relational Databases and SQL

A database is any organised collection of data. This includes spreadsheets, phone books, printouts organised in folders, and the like. In computer science and especially in Bioinformatics the term database normally refers to collections of data that are managed by *database management system (DBMS)*.



**Figure 1.5:** An exemplary profile. The X-axis specifies the position in the sequence. The Y-axis shows the frequencies of the letters within the graph (amino acids) at a given position within the sequence. At position 7 should be either amino acid F or amino acid C. F and C do not sum up four bits (100%) because any other amino acid may occur at position 7 as well, it is “just” unlikely.

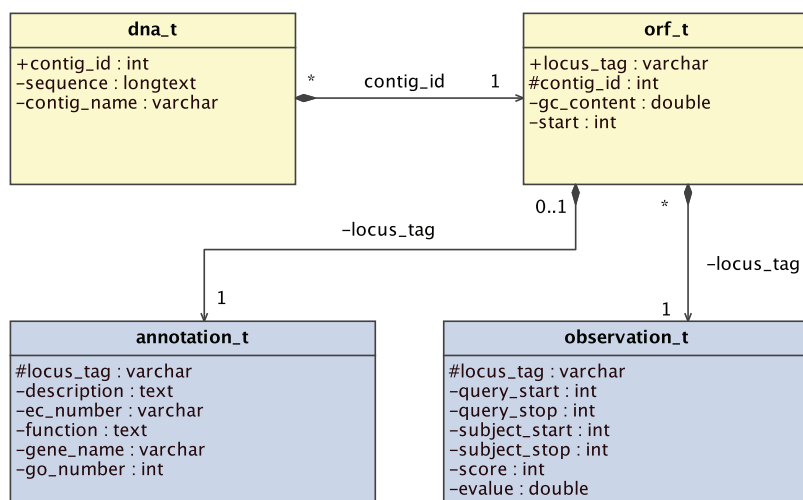
Unlike the afore mentioned ‘database types’, a database management system organises the data; it assures the syntactic correctness, and to a certain degree, depending on the used management system, the semantic correctness. Databases managed by DBMS are commonly accessed using a standardised language, the *structured query language (SQL)*. This language defines standard means: to define database schemes, to insert data, to update data, and to retrieve data. One of the most important features of a database management system, however, is the assurance of data consistency while the data is concurrently accessed.

The most commonly used type of database management system in Bioinformatics is the *relational database management system (RDBMS)*. It uses tables and relations between the tables to organise data. The goal is to reduce redundancy in multiple rows of the same table by splitting that table in two or more tables, accordingly (*normalisation*). Primary / foreign key constraints are used to link the data in the two tables.

Primary Keys are dates that uniquely identify single rows in a table. In secondary tables, the primary key is referenced by *foreign keys*. Additional constraints, such as *on delete cascade* and *on update cascade*, might be put on this relation. These relations specify the behaviour when the row denoted by the primary is deleted or changed. In case it is deleted, all rows referencing it in secondary tables will also be deleted. On update cascade specifies, that when the primary key is updated it will also be updated in all secondary tables.

A simple example is the separation between the description of an ORF and additional data about that ORF. The *orf.t* table may hold the ORF’s locus tag, and its start / stop position within a complete genome or contig. The locus tag is used as primary key as it uniquely identifies each ORF.

A second table *observation.t* may contain information reported by the BLAST



**Figure 1.6:** A simple database schema that can be used to model the data in genome annotation projects. Table *dna\_t* holds DNA sequences, table *orf\_t* describes an ORF, table *observation\_t* stores information about an ORF as reported by BLAST, and information derived from data in table *observation\_t* are linked in table *annotation\_t*.

In a typical genome annotation project table *dna\_t* should contain only one sequence, the closed genome, or a small number of large genome fragments (contigs). Each entry has a unique numeric id (primary key). The ORFs in table *orf\_t* are linked to the DNA sequence, they were predicted on, by the sequence's numeric id.

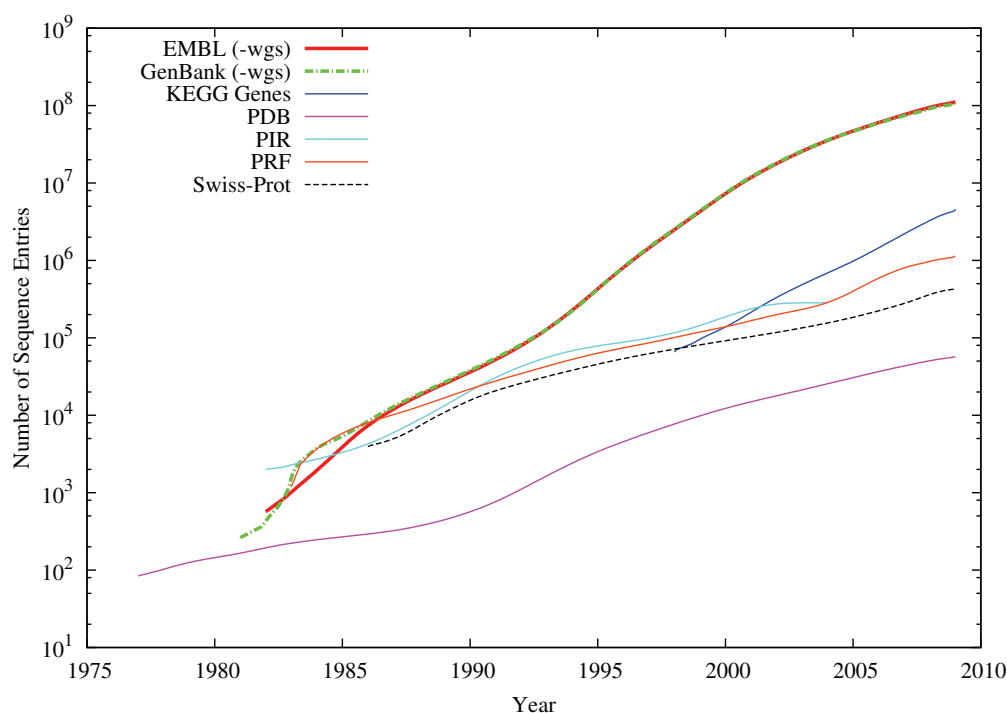
The *orf\_t* table is associated to the tables *observation\_t* and *annotation\_t*. Each entry in table *orf\_t* may be linked to any number of entries in the *observation\_t* table and zero or one entry in table *annotation\_t*. Entries in table *orf\_t* may only exist as long as the corresponding entry in table *dna\_t* exist. If an entry in table *orf\_t* is deleted, then all entries in the tables *observation\_t* and *annotation\_t* must be deleted as well.

tool (31). This information includes: the start / stop positions of the match in both the ORF and the target sequence, the unique ID of the target sequence within the database used by BLAST, the functional description of the target sequence, and the quality values as reported by BLAST. To link the data in the two tables, the locus tag would additionally be added to the *observation\_t* table as foreign key. For each ORF an unspecified number of observation may be reported and, accordingly, an unspecified number of rows in the *observation\_t* table may exist. The separation of data into two table reduces the redundancy of data stored in each table because data describing the ORF will only be held once.

Figure 1.6 depicts a simple database schema that models data produced during genome annotation.

### 1.3.2 Sequence Databases

All DNA sequences, from single protein sequences over genomes to the complete DNA of environmental samples that are described in publications need to be made publicly available. The collaboration (INSDC) of the providers of the databases



**Figure 1.7:** *Growth of Sequence and 3D Structure Databases.*

Source: [http://www.genome.jp/en/db\\\_growth.html](http://www.genome.jp/en/db\_growth.html).

DDBJ (44), EMBL (45), and GenBank (46) provide the resources to publish these sequences. All three databases are synchronised daily. In order to handle millions of entries and the exponential growth of sequence data, these databases cannot be curated. As of June 2009, these databases contain more than 160 million entries comprising more than 275 billion nucleotides<sup>2</sup>. Figure 1.7 shows the increase of publicly available sequences since 1980.

Nucleotide sequence databases are the *primary databases* used for any kind of data mining. *Secondary databases* such as the translated EMBL (EMBL), the non-redundant NCBI nr, and the Swiss-Prot (47) databases provide translations of the protein coding sequences (CDS) found in primary databases. The TrEMBL and the NCBI nr databases are automatically created. Swiss-Prot is a manually curated database that contains only a fraction of the proteins found in primary databases. Additionally it contains protein sequences found in literature which are not contained in the nucleotide sequence databases.

**NCBI nr** is the most widely used database used for the functional description of newly sequenced sequences. It is provided by the National Center for Biotechnology Information (NCBI) and comprises all protein sequences found in the INSDC databases. It also contains sequences from protein sequence databases including the Protein Research Foundation (PRF) database, the Protein Identification Resource (PIR) database (48), the RCSB Protein Data Bank (PDB) (49), the NCBI RefSeq database (50), and the Swiss-Prot database.

<sup>2</sup>source: release 100 of the EMBL database [http://www.ebi.ac.uk/embl/Documentation/Release\\\_notes/current/relnotes.html](http://www.ebi.ac.uk/embl/Documentation/Release\_notes/current/relnotes.html)

The NCBI nr database is non-redundant which means that entries in the source databases that describe the same sequence are merged. The description of the NCBI nr entry contains the descriptions of all merged entries, separated by the merged entry's unique identifier of its source database. An example of an NCBI nr entry is shown in Figure 1.8.

```
putative membrane protein [Yersinia pestis C092] gi|45440854|ref|NP_992393.1|
putative membrane protein [Yersinia pestis biovar Medievalis str. 91001]
gi|22126919|ref|NP_670342.1| hypothetical protein y3042 [Yersinia pestis KIM]
gi|51595516|ref|YP_069707.1| putative membrane protein [Yersinia
pseudotuberculosis IP 32953] gi|21959957|gb|AAM86593.1| hypothetical [Yersinia
pestis KIM] gi|45435712|gb|AAS61270.1| putative membrane protein [Yersinia
pestis biovar Medievalis str. 91001] gi|51588798|emb|CAH20412.1| putative
membrane protein [Yersinia pseudotuberculosis IP 32953]
gi|15979204|emb|CAC89982.1| putative membrane protein [Yersinia pestis C092]
gi|25510076|pir||AC0140 probable membrane protein YP01140 [imported] -
Yersinia pestis (strain C092)
```

**Figure 1.8:** Example of the description of an entry in the NCBI nr database (*gi|16121437|ref|NP\_404750.1|*).

**Swiss-Prot** is a high quality resource for manually annotated protein sequences (47). It constitutes one of the most reliable resources for the functional annotation of proteins available today. Protein sequences from three sources are comprised by Swiss-Prot: the protein sequence database PIR, a subset of entries contained in the TrEMBL database, and sequences from literature. High quality of the annotations, minimal redundancy, and integration with other databases are three criteria by which Swiss-Prot distinguishes itself from other protein sequences databases. Of which the quality of annotations is the most important criteria.

Each sequence entry is manually curated and revised by an expert for the protein family. Single entries or a group of entries (of the same protein family) are periodically updated if new information becomes available. The Swiss-Prot team reduces redundancy in the database by merging separate entries of the same sequence found in the source databases. Swiss-Prot entries contain cross-references to external databases which provide further information. At present, more than 100 external databases are cross-referenced by Swiss-Prot<sup>3</sup>.

While the primary databases and automatically created protein databases grow at an exponential rate, the growth of Swiss-Prot is hampered by intensive manual labor which is invested in the curation process. Among other databases, Swiss-Prot and TrEMBL are now integrated by the UniProt database (51). Within this collaboration the Swiss-Prot database is called UniProt knowledge base (*UniProtKB*). This denotes the high quality of the Swiss-Prot database.

### 1.3.3 Pattern / Profile Databases

Pattern and Profile databases are *secondary databases* which are created from protein databases. They contain patterns or profiles of a group of functionally related proteins or sub regions of a protein that itself constitutes a functional building block (domain). Functionally related proteins are identified in the source

<sup>3</sup>Source: <http://www.expasy.ch/cgi-bin/lists?dbxref.txt>

databases and a multiple sequence alignment is created. A pattern or profile that describes conserved regions in the MSA is created. These patterns and profiles are then used to provide evidences for the functional characterisation of newly predicted ORFs.

**InterPro** is an integrative database that integrates the information provided by eleven independent pattern and profile databases (11). It also includes information provided by the UniProt protein database. Among these databases are the profile databases Pfam (52), and TIGRfams (53). InterPro entries are comprehensive and they reference all entries found in its member databases that describe the same protein family or domain. Extensive cross-references to the referenced entries and to external sources are provided. Each InterPro entry is also classified according to the *Gene Ontology (GO)* and *Enzyme Commission (EC)* classification schemes.

InterPro provides information about the relationship between its database entries. An entry describing a protein family may also belong to a group of proteins that describe a broader function. It is the *child of* another entry. The broader protein family is the *parent of* the more specific entry. An example is the entry IPR000025 which describes the *Melatonin receptor* protein family. This protein family describes a function that is more specific than the function described by the entry IPR000276 (*7TM GPCR, rhodopsin-like*). Hence, the proteins comprised by entry IPR000025 also belong to the group of proteins described by entry IPR000276.

Domains are the building blocks of a protein function. As such an entry describing a domain might be *found in* one or more entries describing a protein family. On the other hand, a protein family *contains* domains. The domain *ADAM, cysteine-rich* (IPR006586) which can be found in the protein family *Peptidase M12B, ADAM-TS1* (IPR013274) is an example of this relationship.

In this work only observations are used that report similarities to entries of the Pfam and the TIGRfams member databases.

**Pfam** contains profiles of protein families and domains based on hidden Markov Models (52) It is divided into two sections: Pfam-A and Pfam-B. Pfam-A is a high quality, manually curated database. It contains the profiles of more than 10,300 protein families and domains. The domain profiles cover more than 74% of the proteins found in the UniProtKB (Swiss-Prot) protein database<sup>4</sup>.

Pfam-B is a collection of profiles derived from automatically created multiple sequence alignments based on entries of the PRODOM database (54). Profiles in Pfam-B do not overlap with profiles found in Pfam-A. It is lower quality than Pfam-A because it is based on automatically created multiple sequence alignments. Pfam-B supplements Pfam-A and covers an additional 19% of the proteins in the UniProtKB database.

**TIGRFAMs** contains profiles of protein families based on hidden Markov Models (53). These protein families are manually curated. A decisive feature

---

<sup>4</sup>Jaina Mistry, Penny Coghill, Sean Eddy, Rob Finn, John Tate and Alex Bateman. Broadening Pfam Protein Sequence Annotations. Available from Nature Precedings <<http://dx.doi.org/10.1038/npre.2009.3194.1>> (2009)



of the TIGRFAMS database is the classification of *equivalogs*. While protein families found in other database might described a group of paralogous proteins that do not share the same function, equivalogs describe a group of proteins which necessarily share the same function.

Observations based on the these manually curated InterPro member databases constitute high quality evidences for the functional annotation of proteins.

### 1.3.4 rRNA Databases

Projects that provide databases of the SSU rRNA marker gene include: the Ribosomal Database Project (RDP) (55; 56), the European ribosomal RNA database (2), the Greengenes project (GG) (3), and the manually curated ARB databases curated by Dr. Wolfgang Ludwig (Department of Microbiology – TUM). All databases are automatically created, except those provided by Dr. Wolfgang Ludwig.

The European ribosomal RNA database has been discontinued due to funding problems.

The RDP project focuses on bacterial and archaeal SSU rRNA sequences only. As of release 10, it uses the *Infernal* alignment software (57). Also, this project does not provide the alignment in the ARB database format which makes it difficult to be used in combination with the ARB software suite. As of release 10 update 13 (July 28, 2009) the database provided by RDP contains 1,049,433 automatically aligned SSU rRNA genes.

The database provided by the Greengenes project also only include SSU rRNA sequences from the bacterial and archaeal domains. Compared to the RDP database, it only covers full length sequences<sup>5</sup>. The alignment of the sequences is created by the NAST aligner software (58). Besides other formats it provides sequences in an older version of the ARB database format. 397,006 are included in the released database since June 26, 2009.

ARB databases are manually created and offer high quality alignments of sequences of all three domains of life (*Archaea*, *Bacteria*, and *Eukarya*). Due to the manual curation of the alignment and the exponential increase of available sequence data they are limited in content (approximately 40,000 SSU sequences – last officially released in January 2004 updated in February 2005).

## 1.4 Sequencing Artifacts

### 1.4.1 Vectors and Vector Contamination

Vectors are short circular stretches of DNA that are able to replicate independently of the chromosome. In molecular biology, vectors are used among other things to clone certain pieces of DNA as a prerequisite to sequencing. A single gene of interest, obtained through PCR amplification, is inserted into the vector, which itself is then introduced into the cell of an organism that can be easily grown, e.g. *E. coli*. During the reproduction of the cells, the DNA fragment carried by the vector will also be amplified. After sequencing, the vector sequence information is cut off in silico and the relevant sequence information is extracted.

---

<sup>5</sup>sequences longer than 1250 nucleotides

In cases where this is not done or where the vector sequence and the start of the DNA fragment can not be distinguished unambiguously, this leads to vector contamination.

### 1.4.2 Chimeras and Chimera Detection

Chimeras are sequences artificially created during PCR based DNA amplification that are composed of parts of two or more individual sequences. These sequences may or may not belong to the same organism but they must be related (e.g. encode for the same gene).

Pintail (59) is a software tool used to detect sequence anomalies and can also be used to identify chimeric sequences. In its downloadable form, it provides a graphical user interface to check a single 16S rRNA query sequences. It aligns the query sequence and closely related sequences to a references 16S rRNA sequence of *Escherichia coli* using clustalw (40). The most likely *break point* is then reported to the user and it is left to the user to interpret the results. The break point is the nucleotide position within the query sequences where two sequences are most likely joined.

The Pintail software is released as GPL software and the RDP project modified its source to be better suited for batch processing in large scale projects. This version of Pintail uses a FASTA file as input. The input file must contain an even number of sequences. Every odd numbered sequence is a query sequence and the following even numbered sequence is a closely related sequence. Like in the standard version of Pintail, a multiple sequence alignment of these sequences and the *E. coli* reference sequence is created. Unlike the standard version, the modified version reports the results for each tuple of query and subject sequence on the command line. For each query sequence it reports the sequence identifier, the most likely break point, the expectation value of the break point, and it draws a conclusion. It uses the values *yes*, *no*, *likely*, *unknown* to denote if the sequence is found to be *a chimera*, *no chimera*, *a likely chimera*, or that the software could not decide *cannot tell*.

Pintail uses a hard-coded the 16S rRNA of *E. coli* as references sequence and is trained on a 16S rRNA dataset. It can therefore not be used to reliably check other sequences than 16S rRNA sequences. A substitution that can also be applied to check 18S rRNA sequences as well as LSU rRNA sequences is currently developed by Karin Dietrich in Microbial Genomics Group at the Max Planck Institute for Marine Microbiology in Bremen, Germany.

The modified version of Pintail is used by the RDP project to prune chimeric sequences from their database. The Greengenes project uses the Bellerophon software tool for the detection of chimeric sequences (60).

## Chapter 2

# Research Objectives

Advancements in sequencing technologies and the growing interest in Metagenomics, has lead to an unprecedented mass of available sequence data. The amount of available sequence data continues to grow exponentially. To deal with such huge amounts of sequence data, the field of biology has to leave the days of craftsmanship behind which produces primarily unique results. It has to enter the era of industrialisation making results reproducible and making methods applicable in large scale studies.

The objectives of this work are therefore to provide biologists automatised tools to shape the continuous flow of sequence data. These tools should enable biologists to gain insight into whole microbial community structures as well as their functional diversity.

### 2.1 Phylogenetic Classification

The goal of this thesis is to create a system to automatically provide databases of high quality alignments comprising all publicly available SSU and LSU rRNA sequences. Tasks that need to be solved to create such a system include:

- automatic retrieval of candidate sequences from public databases,
- quality assessment of imported sequences,
- fully automated alignment of candidate sequences,
- providing pre-configured databases, and
- providing a web interface for easy data access.

In addition to keyword based sequence retrieval, sequences should also be retrieved based on sequences similarity. The quality assessment should include the quantification of ambiguous bases, homo polymeric stretches, and vector contamination as well as include a chimera check. Pre-configured databases should be available in the ARB database format (amongst others). A web presence is to be established to enable users to create custom databases. Additionally, tools to inspect, to search in, and to align novel sequences against the databases should be made available.

## 2.2 Functional Characterisation

In (10) a prototype for the automatic annotation of genes was implemented using the horizontal annotation approach. Evaluation of this prototype showed good performance compared to the manual annotation of the organism *Gramella forsetii* KT0803 (61). Especially for ORFs without a functional prediction the computer outperformed the human annotator in terms of consistency and reproducibility. Although functional, the prototype did not include InterPro (11) domain predictions and only limited assessment of the reliability of functional predictions.

In this thesis the prototype is to be extended to include InterPro domain prediction to characterise ORFs without a functional assignment. Functional annotations need to be labelled according to their reliability to be able to easily screen for problematic annotations. Additional performance studies and a proper comparative evaluation of the system need to be made especially when compared to tools using the vertical annotation approach.

## Chapter 3

# Summary

The work of this thesis covers two distinct topics within the field of Bioinformatics (focused on environmental microbiology): the development of a pipeline to automatically create databases for the phylogenetic classification of organisms based on selected marker genes, and the development of a software tool for their functional characterisation based on genomic data. Additionally, a contribution was made to the initial development of a database which associates genomic data with environmental and contextual data.

In the SILVA project, a pipeline was developed to automatically create comprehensive databases of the SSU and LSU marker genes including sequences from all three taxonomic domains as well as non full-length sequences. This pipeline includes tasks to: extract annotated sequences, predict rRNA in otherwise not annotated environmental samples, import mis-annotated sequences based on white list filters, check the quality of the imported sequences, align sequences, import additional meta data from third parties, and export the whole database or parts in various formats, including the ARB database format.

In the second part of this thesis, the enhancement of the MicHanThi software tool developed in (10) are addressed. This software tool is used to automatically create annotations for genes based on the results of homology search tools like BLAST (31), InterProScan (8), SignalP (62), and TMHMM (63). As part of this work, MicHanThi has been extended to include InterPro (11) domain predictions as well as to consider relations among InterPro database entries in the prediction of gene functions. Additionally, an evaluation of the results (annotations) created by MicHanThi as compared to those created by the RAST web service was done.

The last part of this thesis is concerned with the development of a database to link genomic sequence information with environmental and contextual data. Marine environments present the focus of the project and the database mainly includes information relevant for these habitats. Information included are: the depth a sample was taken, the water temperature, the pressure, and the salinity among many others. Web based tools were developed to provide an interface to query the database (<http://www.megx.net>). To answer more complex questions GUI applications such as Metalook (64) and Metamine (65) were developed.

### 3.1 SILVA

An important aspect of molecular microbial ecology is to understand the structure of microbial communities and to identify and quantify the structural composition of organisms. Comprehensive databases of aligned sequences and tools to handle hundreds of thousands of sequences in these databases are essential for this task. ARB (from Latin arbor – tree) is a widely used software workbench which is used for phylogenetic reconstruction.

Until 2004 the main databases for this program were provided by Dr. Wolfgang Ludwig (Department of Microbiology – TUM), including SSU and LSU rRNA genes. These databases were manually curated. Due to the exponential growth in the number of sequences in publicly available databases, these databases can no longer be considered comprehensive. Three projects (see Section 1.3.4 for details) try to provide up-to-date and comprehensive databases of automatically aligned sequences. Limitations of these databases include that no database contains sequences from all three domains and sequences with a length of less than one thousand nucleotides. A second important drawback is that only the Greengenes project provides databases in the ARB database format. As an additional limitation, the format chosen for the alignment is based on the ARB alignment of 1997 which constitutes approximately 7,000 positions. Currently, ARB alignments are 50,000 positions wide. The SILVA project (from Latin – forest) was initiated, to overcome these limitations. Its focus is to provide automatically created, comprehensive, quality controlled, meta data enriched, and ARB compatible databases of publicly available LSU and SSU rRNA genes.

The SILVA project provides two distinct quality controlled databases: one for the small subunit rRNA gene (SSU) and one for the large subunit rRNA gene (LSU) – the *Parc databases*. These databases include all aligned rRNA sequences that suffice certain quality standards. For each database a smaller subset of high quality, full-length sequences<sup>1</sup> is made available – the *ref databases*. As of release 100 (August 2009) of the SILVA databases, the Parc databases include 995,747 SSU and 161,017 LSU sequences (Ref databases: 409,907 SSU and 14,426 LSU sequences).

Custom tailored databases that include a more focused subset of sequences, e.g. only sequences of a certain phylogenetic group, can be created by users on the SILVA web site at <http://www.arb-silva.de>. All databases are available in the ARB database format as well as in the (aligned) FASTA format. Various custom formats are provided to incorporate information contained within SILVA into other projects. For example, the EMBL database includes a reference to SILVA for all rRNA sequences in the SILVA Parc databases.

Lately, the SILVA pipeline and its tools are not only used to build the SILVA Parc and Ref databases, but it is also used to analyse in-house 454-pyrosequencing data sets, e.g. samples from the Logatchev site (*Regina Schauer unpublished*; Department of Molecular Ecology – Max Planck Institute for Marine Microbiology).

In these projects, all 454-pyrosequencing reads are imported into a custom database, regardless of their genomic information. Sequence checks are applied to reject bad sequences. Datasets obtained by 454-pyrosequencing can not contain

---

<sup>1</sup>For the SSU database: 1,200 nucleotides (*Bacteria, Eukarya*), 900 nucleotides (*Archaea*). For the LSU database: 1900 (*Archaea, Bacteria, Eukarya*)

sequences contaminated by vectors, as a cloning step is not included. Therefore, the vector check is omitted. All remaining sequences are then aligned and exported to an ARB file.

From the 1.2 million reads of the Logatchev sample, 1,385 LSU sequences and 702 SSU sequences could be identified. These sequences were then added to the guide tree of the LSU / SSU Ref databases and an estimate of the diversity within the sample could be gained.

These are interim results and their validity must be established. However, this example shows that the concept of the SILVA pipeline, even though implemented for a very specific task, offers enough flexibility to use parts of it or the whole pipeline for tasks it was not initially thought to solve.

### 3.1.1 Tasks

SILVA is divided into several tasks that are executed after one another: the importer, the quality management, the aligner, and the exporter. A central relational database is used as persistent storage and each task modifies the data in this database as necessary. The workflow in the SILVA pipeline and the interactions between the different tasks are shown in Figure 3.1.

Multiple instances of each task can be run in parallel and each instance works independently on a subset of all sequences in the database. A BASH script is used to initialise the different tasks and the Sun Grid Engine (SGE) is used to distribute the tasks on a compute cluster. The dependencies, when can a task be executed, are managed by the SGE's *-hold\_jid* option.

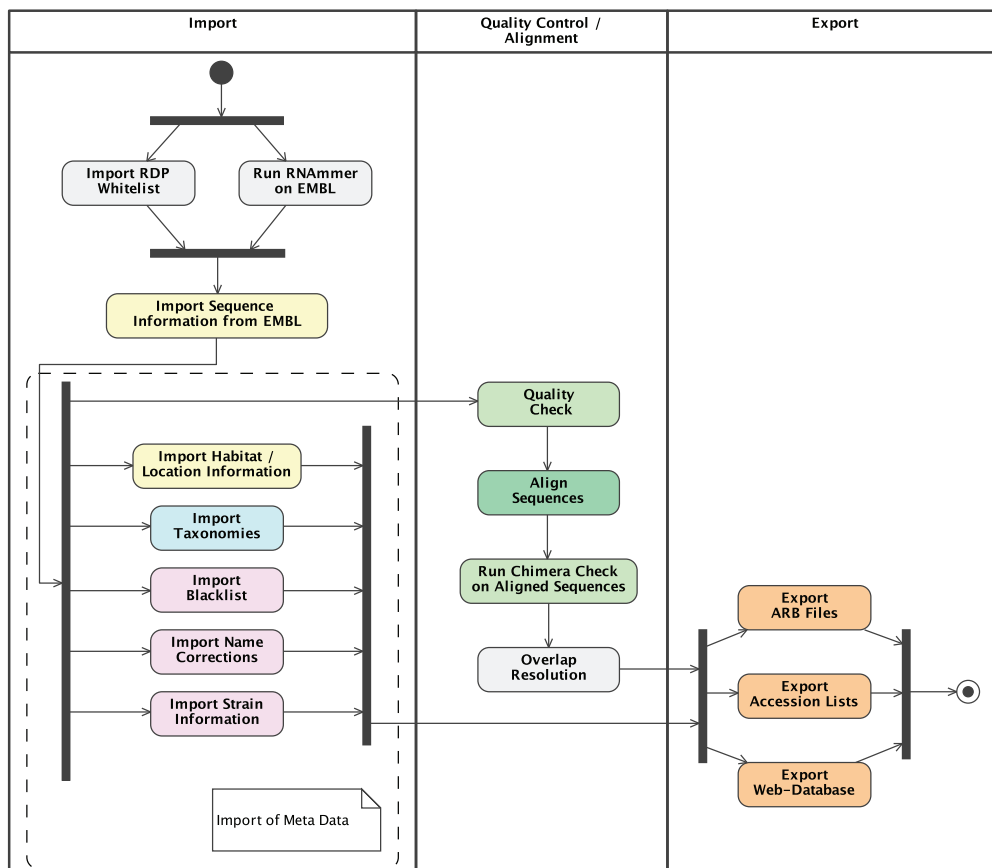
For each task, a job to initialise that task is submitted to the grid engine. All tasks, except the import task, are set on hold and, therefore, won't be executed before the previous task has been finished. The 'initialise jobs' are used to create *SGE job arrays* for the actual task based on the number of sequences in the database and the time a task needs to process a single sequence. This job also manipulates the hold flag for the following 'initialise jobs' so that these jobs now wait for the submitted array job.

**Sequence Import** INSDC is the main source for rRNA sequences, specifically, the EMBL database. All sequences marked as SSU / LSU rRNA gene and sequences marked as 'generic' rRNA gene<sup>2</sup> are imported into the SILVA Parc databases. When the list of imported SSU rRNA genes was compared to the list of sequences included in the RDP database it was shown that the importer missed sequences. Inspecting the EMBL entries of missing sequences revealed that these sequences were wrongly classified in EMBL or that these sequences were not annotated at all. To include these sequences in SILVA databases, a whitelist containing all entries found in the RDP database is used. Additionally, all *misc\_rRNA* genes are checked and the RNAmmer tool (69) is used to scan all EMBL entries for otherwise not annotated SSU / LSU rRNA genes.

Applying only relaxed criteria for the import of sequences and using a whitelist as well as scanning for not annotated rRNA sequences, ensures a high sensitivity of the SILVA databases. The specificity of the SILVA databases is ensured in a later step by the aligner.

---

<sup>2</sup>unspecific rRNA annotations which do not state the type of rRNA



**Figure 3.1:** Workflow and interactions in the SILVA pipeline. The SILVA pipeline can be separated into three main task groups: import tasks, ‘worker’ tasks (quality control and alignment), and export tasks. The import tasks can be subdivided into tasks that import primary data and those that import (third party) meta data.

Tasks in the box titled ‘Import of Meta Data’ are semantically grouped. A list of habitat and location information is internally curated by the *mexr* project (66). These information are imported by the task ‘Import Habitat / Location Information’. The tasks grouped in ‘Import Taxonomies’ import third party taxonomies provided by the Greengenes (3) and RDP (56) projects. A blacklist of organisms not to be included in the exports is internally curated and it is augmented by a list of accession numbers provided by EMBL (67). It is imported by the task ‘Import Blacklist’. Corrections to the names of organisms are provided by the DSMZ (<http://www.dsmz.de/download/bactnom/names.txt>) and by the Livingtree project (68) (tasks ‘Import Name Corrections’). Type strain, strain, and genome information are provided by EMBL the Livingtree project, the RDP project, and the StrainInfo project and are imported by tasks in the group ‘Import Strain Information’.

The colour used for each task denotes the colour of the table, which is modified by the task, used in the SILVA database figure (Figure 3.3).



**Quality Management** All sequences imported into the SILVA databases are checked: for ambiguous bases, for repetitive nucleotides (homopolymers), for vector contamination, and for being chimeras.

Ambiguous bases are nucleotides that could not be unambiguously resolved during sequencing. Homopolymers are parts of a sequence where the same nucleotide occurs at least four times in a row. This is commonly caused by sequencing robots when the same base is sequenced multiple times. Vector contamination is caused by an inaccurate post processing of sequences by sequencing software leaving parts of the vector attached to either end of a sequence. Chimeras are sequences artificially created during PCR based DNA amplification that are composed of parts of two or more individual sequences.

A sequence may have at most 2% of ambiguous bases, 2% of homopolymers, and 5% percent of vector contamination. Additionally, it must be at least three hundred nucleotides in length. These thresholds are based on the quality distribution of the sequences in the initially released SILVA databases. A sequence will not be aligned if it fails one of the quality thresholds. Instead, it will be marked based on its failing quality attribute.

If the sequences is less then three hundred bases long or if one of the thresholds exceeded then its quality is always 0, otherwise its quality is calculated as:

$$Quality_{seq.} = 1 - \frac{\frac{\% \text{ ambig.}}{\text{allowed}(\% \text{ ambig.})} + \frac{\% \text{ homop.}}{\text{allowed}(\% \text{ homop.})} + \frac{\% \text{ vector}}{\text{allowed}(\% \text{ vector})}}{3} * 100 \quad (3.1)$$

In SILVA, BLAST (31) is used to compare each sequence to a database of known vectors. This database comprises the publicly available vector databases of NCBI and EMBL. All rRNA vectors were removed from the final database to avoid miss classification. The longest (inclusive) stretch of vector contamination at either end of the sequence is calculated from all matches against the vector database for a particular sequence. These stretches are then compared to the length of the sequence and the percentage of vector contamination is reported.

The result of the chimera check is not considered in the overall sequence quality as it is does not describe the intrinsic quality of a sequence. Also, it can not be quantified: a sequence is either a chimera or it is not.

To classify a sequence as chimera, SILVA compares each small subunit rRNA query sequences to up to ten closely related sequences using Pintail (69). A chimera-free subset of the SEED and the ARB positional tree (PT) server are used to find close relatives for the chimera check. The ARB PT-server uses a suffix tree to index all sequences in the SEED. The complexity of searches using the PT server is logarithmic, which is an advantage over the tools like BLAST whose searches are linear (1)

From the results reported by the software the *pintail quality* value is calculated:

$$Quality_{pintail} = \frac{\text{number}(\text{negative tests}) * 2 + \text{number}(\text{likely tests})}{\text{number}(\text{tests done}) * 2} * 100 \quad (3.2)$$

The result of the chimera check is presented as additional quality value and it is left to the user to decide how to handle sequences considered to be chimeras.

Large subunit rRNA sequences can not be checked at present because the Pintail software is not able to handle such sequences (69). A replacement software that works independent of the type of input sequence is currently developed by Karin Dietrich at the MPI in Bremen.

**Alignment** Each sequence that suffice the minimal quality standards (sequence quality  $> 0$ ) is aligned against pre-aligned, closely related sequences. The multiple sequence alignments used as reference alignments (SEEDs) for the alignment of new sequences is based on the last release of the manually curated ARB databases released in January 2004. It has been extended by Dr. Katrin Knittel (Department of Molecular Ecology – Max Planck Institute Bremen) to also include sequences from the *Archaea* domain. Additionally, the alignment was curated by Prof. Dr. Frank Oliver Glöckner and student co-workers. These databases provide a high quality alignment.

SILVA uses a custom program for the alignment of sequences: the SILVA INcremental Aligner (SINA), written by Elmar Prüße (Microbial Genomics Group – Max Planck Institute Bremen) as part of his diploma thesis (20). This program uses the alignment of sequences in the SEED as a reference for the alignment of a new sequence to create an incremental alignment. It does not create a de novo multiple sequences alignment because the SEED alignment has been manually created for more than fifteen years. Also the number of sequences is too large to create a new multiple sequence alignment from scratch.

For each query sequence, five to forty closely related sequences are searched for using the ARB PT server. Among these sequences at least one sequence must be a full-length sequence. It must be longer than 1200 nucleotides for eukaryotic and bacterial SSU sequences, or it must be longer than 900 nucleotides for archaeal SSU sequences. For the LSU database, sequences are considered to full-length if they are at least 1900 bases long (all three domains). The reported sequences serve as reference for the alignment of the new sequence. The aligner starts by aligning the query sequence to the closest relative. If the alignment score drops beneath a defined threshold then the next closest relative is used. This process is repeated until the query sequence is fully aligned.

For each aligned sequence, a number of attributes are reported: the base pair (BP) score, information about the close relatives, and the alignment quality. The BP score estimates the stability of the folded molecule based on secondary structure information. The alignment quality is an estimation of the probability that the sequence is correctly aligned. It is based on the alignment's normalised score which represents the alignment distance of the new sequence to selected set of sequences.

The aligner assures the specificity of the SILVA databases. A detailed explanation of the alignment process can be found in (20).

**Meta Data** The complete header section of an EMBL entry is parsed and imported into SILVA. A selected number of feature qualifiers defined in an entry's source feature are also imported. The official INSDC taxonomy of each entry is imported as part of the EMBL header. Additionally, the taxonomic classifications as defined by the Greengenes project and the Ribosomal Database Project are imported. If an entry included in SILVA lacks a classification in on of the projects

then it is assigned to the group of *Unclassified* organisms for this project. For the LSU databases no additional taxonomy is imported since no further up-to-date databases of aligned LSU rRNA sequences exists, at the moment.

Further, information from various third party sources is also imported into the SILVA database. This information includes: culture and type strain information, environmental information, updated and corrected organism names. Culture and type strain information are imported from information provided by the Living Tree project (68), RDP, StrainInfo (70), and EMBL. Each entry in SILVA is marked accordingly. The character ‘T’ denotes organisms that form a type strain. ‘C’ is used to label cultivable organisms which also include type strains. Additionally, the markers ‘l’, ‘r’, ‘s’, and ‘e’ are assigned to each attribute to document the source of the information.

The German Collection of Microorganisms and Cell Cultures – Deutsche Sammlung für Mikroorganismen und Zellkulturen (DSMZ) – curates a list of changes made to the nomenclature of organism<sup>3</sup> based on the official information released by the ‘International Journal of Systematic and Evolutionary Microbiology’. This list is updated monthly. It provides corrections of the spelling of organism names as well as updated names. In SILVA, the official organism name as provided by EMBL is substituted by this information. The old name of the organism is kept. If an organism has been renamed multiple times then the complete history of changes to the name is recorded in SILVA.

A complete list of third party information and EMBL feature qualifiers imported into SILVA and the source of the information can be found in Appendix B (Table B.2).

**Data Export** The information included in SILVA is provided in various standard and custom formats. The three main exports are: the Parc database, the Ref databases and the web database. The SILVA Parc and Ref databases are provided in the ARB database format as well as the FASTA format (aligned sequences with comments). The web database is a reduced version of the SILVA SQL databases only containing sequences also contained in the Parc database. It is optimised for queries executed by the web site.

Further exports are: a list of primary INSDC accession numbers, a list of primary accession numbers associated with quality values and a direct link to access the sequences in a web browser, and a list of primary accession numbers with the different taxonomic classification contained in SILVA for each sequence.

### 3.1.2 Web Presence

Pixelmotor<sup>4</sup>, a company for web design, was contracted to design and implement the initial version of the SILVA web presence. Since then, the web site has been rewritten in most parts keeping the original design. In the last two years, new sections have been added to the web site making SILVA more than just a provider of aligned sequence databases. Especially the sections about *fluorescence in situ hybridization (FISH)* (19), probe design, and the proposed work flow for *Standard Operating Procedure for Phylogenetic Inference (SOPPI)* (71) transformed the

<sup>3</sup>Nomenclature up to date: <http://www.dsmz.de/download/bactnom/names.txt>

<sup>4</sup>Pixelmotor is succeeded by the ANIMA Entertainment GmbH

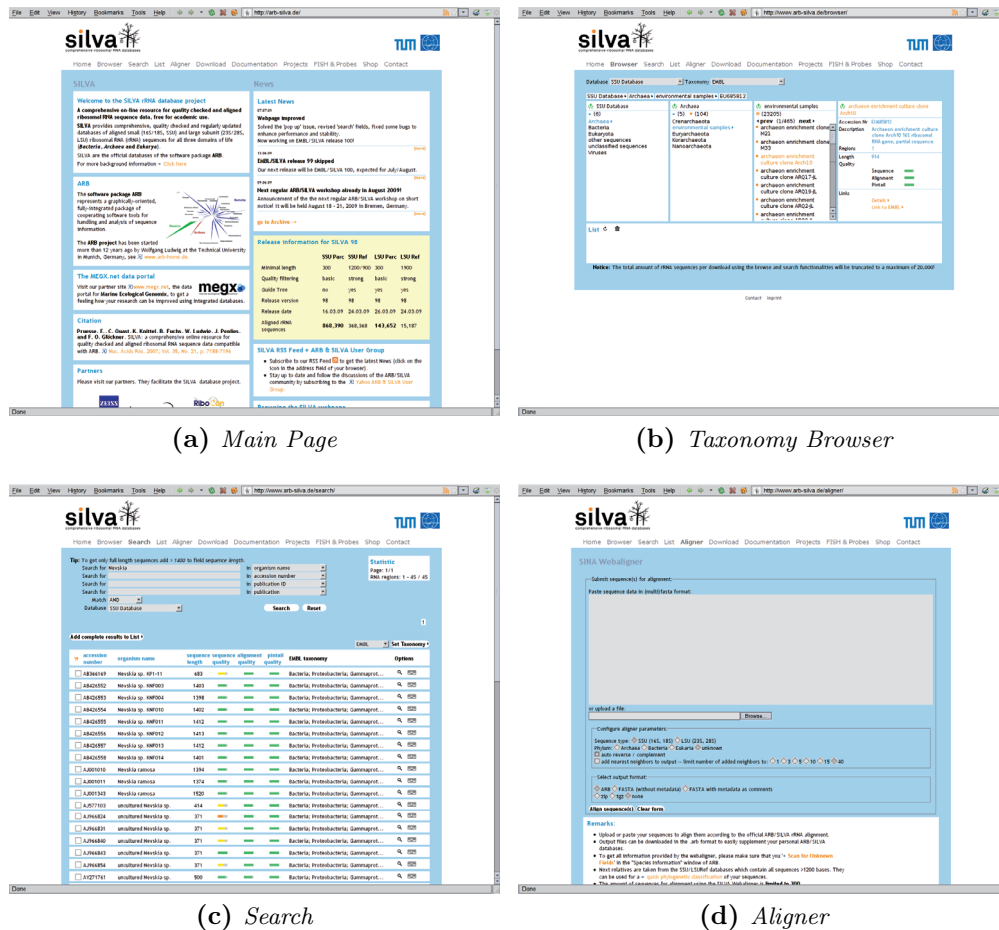


Figure 3.2: The SILVA web presence at <http://www.arb-silva.de>.

SILVA web presence into a one-stop-site for biologists interested in the ecological study of organisms and organism communities. Furthermore, the SILVA projects hosts the official web site of the Living Tree project (68).

In its original version, the web site provided common information about the project (Fig. 3.2a), a taxonomy browser (Fig. 3.2b), a ‘Search page’ (Fig. 3.2c), a ‘List page’, and a ‘Download page’. A web front-end for the SINA aligner has been added to complement the functionality of the web site, since then. Information about the SILVA project, the related ARB Project, and statistics about the current releases of the SILVA databases are found on the main page.

The taxonomy browser allows users to browse the SILVA Parc databases. By default it uses the EMBL taxonomy. Additional taxonomies included in SILVA can also be chosen to navigate the databases. At the time of writing these include the Greengenes and RDP taxonomies for the SSU database. For each entry, a pop-up can be accessed which provides detailed information about the organism: the organism name, the primary INSDC accession number, a list of publications, a list of rRNA sequences included in SILVA, the three distinct quality values for the sequence, the chimera check, and the alignment, and selected meta information such as the GPS position.

On the ‘Search page’, organisms can be searched for by organism name, by

INSDC accession number, by publication (ID), by strain, by sequence length, by quality assessment, by taxonomic path, by submission date, and by ‘sequence entry’. The two search fields publication and ‘sequence entry’ are meta fields that include a search in all text fields of the tables Publication and SequenceEntry. Detailed information about the results and a direct link to the taxonomy browser are provided. Results of the search, taxonomic sub groups, as well as single entries selected in the browser can be added to a sequence cart, the ‘List’. Selecting single entries and sub groups allows users to create custom versions of the SILVA databases specially focused on their needs.

The status of this ‘List’ can be accessed through the ‘List page’. On this page, the ‘List’ can also be prepared for download. Either the FASTA format (aligned) or the ARB database format can be chosen as a download format.

The ‘Aligner page’ (Fig. 3.2d) allows users to align their own sequences using the SILVA pipeline. Users can either paste a single sequence into the provided text field or upload a FASTA file containing up to five hundred sequences. Users can optionally choose to include up to forty nearest neighbours in the result.

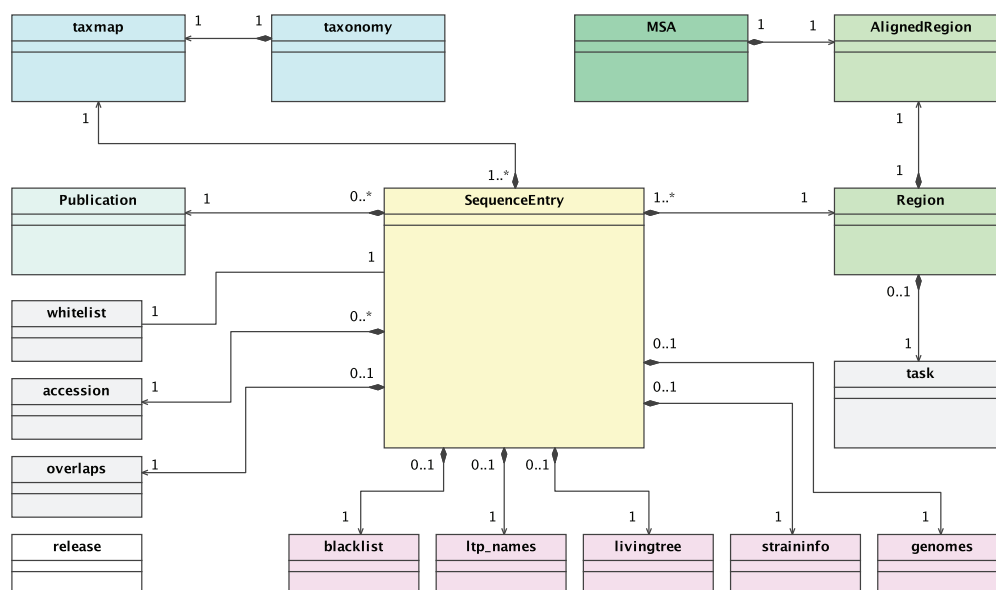
Ready-to-use versions of the SILVA databases can be downloaded from the ‘Download page’. This page also provides links to archives of the ‘List’ prepared for download and a user’s sequences aligned by the SILVA aligner. Users have access only to their own custom databases and aligned sequences.

### 3.1.3 Design and Implementation

**Tools and Pipeline** The SILVA sources are divided into three libraries: the database abstraction library, the IO / tool library, and the aligner library. The database abstraction library provides an in-memory representation of the used data and an interface class that defines an infrastructure to persistently store the data on disk and to load the data from disk. The IO / tool library implements the interface and uses the MySQL relational database management system to store the data. It also implements the importers, the ARB exporter, the sequence check, and the chimera check modules. The aligner library provides the implementation of the aligner. It overlaps with the two other libraries in certain parts because it is designed to also work independently of SILVA.

**Database / Data model** The database as well as its in-memory representation is closely modeled based on the EMBL file format<sup>5</sup>. The central class and table is the *SequenceEntry*. It holds most meta data about an entry that is found in the header section of the EMBL file format: its primary INSDC accession number, a list of secondary accession, the sequence version as specified in the entry, the dates the entry was submitted, imported into EMBL, and when it was last modified. Additional, selected feature qualifiers from the source feature of the feature table section of an EMBL entry are also imported, as well as meta data provided by third parties. See Table B.2 in Appendix B for a complete list of data imported into SILVA databases. Publications which are also part of the header are represented by their own table and class (*Publication*). RRNA sequences described in the *feature table* section of the EMBL format are stored in the *Region* table / class. A region may belong to more than one multiple sequence alignment

<sup>5</sup>[http://www.ebi.ac.uk/embl/Documentation/FT\\_definitions/feature\\_table.html](http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html)



**Figure 3.3:** The design of the SILVA database. The table *SequenceEntry* (yellow) is the central data object which connects the taxonomic information (blue), sequence information (green), and meta data (purple) stored in the SILVA databases. The meta data tables are dynamically created when the associated information is imported and may not exist in all databases. The information contained in these tables is also added to the content of the associated fields in the *SequenceEntry* table. Therefore, these tables are only used to document the changes made to entries in the *SequenceEntry* table. Tables depicted in gray are organizational tables. Their names and the names of the meta data tables are in lower case letters, to further indicate their ‘temporary’ nature.

defined in table / class *MSA*. The alignment of the same sequence may differ between different MSAs, therefore, the table / class *AlignedRegion* was introduced to hold the aligned sequence, information about the alignment reported by the aligner, and a link to the MSA to which a region belongs. External references found in the header, publications, and regions are stored in the *Reference* table / class.

The *one-to-many* relation between *Region* and *MSA* was chosen to be able to easily compare the alignments created by multiple aligner runs with different parameter sets. A second reason is to be able to store the alignments curated by different experts. It was initially planned to store the SEED, used to align new sequences, in the SILVA databases and to provide an interface to extend and to enhance the alignment of the SEED. The one-to-many relation was changed into a *one-to-one* relation because this interface has never been realised and the idea to store the SEED in the database has been dropped. In the current SILVA pipeline the different MSAs are used to differentiate between the possible states of a region in the database.

When a sequence entry is first imported all its regions are assigned to the *MSA imported*. The quality check module then assigns the regions to different MSAs based on their sequence quality, *ambiguous*, *bad length*, *homopolymer* or *vector*. If a region is eligible for alignment then it is assigned to the *MSA unaligned* or

to the MSA *unaligned\_rnammer* if the region was predicted by RNAmmer (69). The aligner will assign a region to the MSA *auto-aligned* if the region could be aligned. Otherwise, it is assigned to *auto-aligned-rejected*. If a sequence could be aligned but the number of aligned bases is below the chosen threshold it is still assigned to the MSA auto-aligned. Those sequences are excluded by the exporter when the data is exported into the different formats. Further MSAs used to mark ‘unwanted’ sequences are: *blacklist*, *ignore*, and *overlaps*. Regions are assigned to the MSA blacklist based on a list of primary accession numbers manually curated by Dr. Wolfgang Ludwig and Prof. Dr. Frank Oliver Glöckner. It also contains accession numbers provided by EMBL. Sequences predicted by RNAmmer that overlap with sequences contained in EMBL are assigned to the MSA overlaps because and are, therefore, ignored.

Taxonomies associated to each entry are stored in table *taxonomy*. A mapping between the taxonomic paths stored in table *taxonomy* and entries stored in table *SequenceEntry* are provided in table *taxmap*. The concept behind these tables is an adapted version of the *path enumeration model* described by Celko in (72). Each entry in tables *taxonomy* and *taxmap* also hold, additionally to the taxonomic information, the name of the taxonomy. Therefore, multiple taxonomies can be stored in the same table. Currently, each entry in the table *SequenceEntry* is associated to the taxonomies of EMBL, Greengenes, and RDP.

The design of the database is depicted in Figure 3.3.

**Website** The web site is implemented using the programming languages HTML, JavaScript, and PHP. It uses the typo3<sup>6</sup> content management system. A content management system allows content providers to easily modify web pages without the need to know details about web programming. For programmers, that work on the server side of a web site, it offers a framework for web site development (typo script). As such, the taxonomy browser, the search page, the cart, the list, and parts of the download page are implemented using this framework.

The web site uses a denormalised version of the SILVA database and merges information from the meta data tables with fields in table *SequenceEntry* that are not used for querying. It only contains sequences that were automatically aligned and suffice the quality standards for the Parc databases. It is identical to the Parc databases. Therefore, the discrimination between regions and aligned regions is not necessary and the two tables are merged. The table *MSA* is additionally no longer need and has been dropped. All Meta data tables are currently purged from the database. These modifications to the database design were made to improve the performance for a read only query pattern.

**Programmin Languages & Build Dependencies** The SILVA build system is based on the GNU Autotools collection: Autoconf<sup>7</sup>, Automake<sup>8</sup>, and libtool<sup>9</sup>. Hence, it follows the classical *./configure && make && make install* approach that numerous open source UNIX projects use. The tool binaries are imple-

<sup>6</sup><http://www.typo3.org>

<sup>7</sup><http://www.gnu.org/software/autoconf/>

<sup>8</sup><http://www.gnu.org/software/automake/>

<sup>9</sup><http://www.gnu.org/software/libtool/>

mented in the C++ programming language, the submit script, used to manage the SILVA pipeline and that is used to submit jobs to the SGE, is implemented in the *Bourne-again shell (BASH)* scripting language. RNAmmer was originally implemented in Perl (69) and it was adapted for the SILVA pipeline by Felix Schelsinger (former student at the Jacobs University Bremen). To be able to use it to scan the complete EMBL database, it has been rewritten in Python to increase performance by Elmar Prüße (Microbial Genomics Group – Max Planck Institute for Marine Microbiology).

The following external C/C++ libraries are required to build the SILVA sources: ARB,<sup>10</sup> libbz2,<sup>11</sup> libmysqlclient,<sup>12</sup> libpcrc / libpcrcpp,<sup>13</sup> libphoenix,<sup>14</sup> and libz.<sup>15</sup> Additionally, the following Boost<sup>16</sup> libraires are required: Filesystem, Program Options, Regex, Serialization, and Thread.

ARB (1) does not provide a development package. Therefore, the ARB sources have to be compiled before SILVA can be build. The option *-with-arbhome* needs to be passed to SILVA's configure script. It has to point to the ARB build tree. The ARB sources are used to natively support the ARB database format, both for reading and writing, as well as to query the ARB PT server (1). libphoenix is part of the Phoenix EMBL parser and provides support for parsing files in the EMBL format. It is dynamically linked against libpcrc. As part of the SILVA project the parser was ported to the Autotools build system and a Debian package has been created. The parser has further been adapted to changes of the EMBL format and to support loading of compressed files in formats supported by libz and libbz2. The *C application programming interface (API)* provided by the MySQL client library, libmysqlclient, is used in the IO module to realise the connection to the MySQL server. The Boost libraries are used in numerous places of the SILVA source code where the functionality provided by the C++ *Standard Template Library (STL)* does not suffice. libbz2 and libz are optional and if present enable compressed file support.

## 3.2 MicHanThi

In (10) the prototype of a software tool (MicHanThi) was developed to automatically propose functions for potential genes based on the results of homology searches (observations). To predict a gene function, observations reported by the BLAST and InterProScan tools are used (31; 8). If a function cannot be assigned to an ORF then the observations reported by SignalP (62) and TMHMM (63) are considered to create a more accurate annotation.

As the database for the pairwise comparison of sequences (BLAST), the non-redundant NCBI nr database was chosen. This database was chosen because it forms the most comprehensive database of publicly available sequences. Among others, it includes the manually curated, high quality Swiss-Prot database (47)

---

<sup>10</sup><http://www.arb-home.de>

<sup>11</sup><http://www.bzip.org/>

<sup>12</sup><http://www.mysql.com>

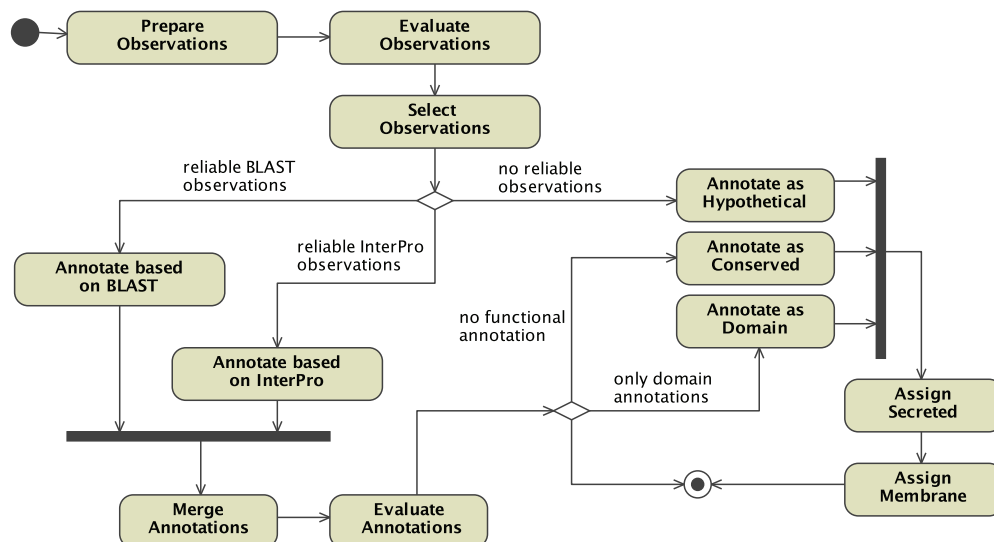
<sup>13</sup><http://www.pcre.org/>

<sup>14</sup><http://www.bioinformatics.org/phoenix/wiki/>

<sup>15</sup><http://www.zlib.net/>

<sup>16</sup><http://www.boost.org>





**Figure 3.4:** The MicHanThi annotation process

which constitutes one of the most reliable sources of functionally annotated proteins. Hence the possibility to functionally annotate as many genes as possible. At the same time it offers the reliable prediction of a gene function if the annotation is based on reliable Swiss-Prot observations.

InterProScan (8) is the second tool used to predict gene functions. It exclusively uses the integrative InterPro database (11). Unlike BLAST, it does not compare a novel sequence to a previously annotated gene. Instead, it compares a novel sequence to the pattern or profile of a group of functionally related proteins (protein family) or to the pattern / profile of a domain found in various proteins. As for annotations created based on reliable Swiss-Prot observations, an ORF may be reliably annotated if the observation reported by InterProScan itself is reliable.

Figure 3.4 depicts the tasks involved by MicHanThi during the annotation. Three main tasks can be distinguished: the processing of observations, the creation of annotations, and the evaluation of annotations.

### 3.2.1 Process Flow

**Observation Evaluation** Based on Fuzzy Logic, observations are assigned to one or more reliability classes (*unreliable*, *uncertain*, *reliable*, *very\_reliable*). The membership functions used to classify observations are based on the thresholds commonly applied by a biologist during the manual annotation of single genes, entire genomes, or metagenomes. For the BLAST tool the characteristics *alignment coverage of the query sequence*, *alignment coverage of the target sequences*, and the *E-value* are considered.

Observations reported by InterProScan describe patterns or profiles. These patterns or profiles may only include a small portion of the protein. Therefore, the coverage of the alignment of neither the query sequence nor the target sequence is decisive for the evaluation of the observations.

SignalP and TMHMM observation report the presence or absence of certain traits. These observations are assigned to classes *uncertain* or *reliable* if the tool reports the presence of the trait. Otherwise the observations are disregarded in the later annotation process.

**unreliable:** Observations reported by both tools with E-value  $> 1e^{-3}$  are considered to be *unreliable*. For BLAST observations, the coverage of alignment in respect to the sequences carries no weight in this case.

**uncertain:** An observation is considered *uncertain* if E-value  $\leq 1e^{-3}$ . For BLAST observations, the alignment coverages must exceed 30% of the length of either sequence. As for the following two classes an E-value closer to zero may balance a low coverage.

SignalP observations are *uncertain* if their probability as reported by SignalP is  $\geq 0.75$  but its cleavage site probability is  $\leq 0.5$ .

**reliable:** Observations showing E-values  $\leq 1e^{-15}$  are commonly assumed to be *reliable*. For BLAST observations, the alignment must exceed 30% of the length of the query sequence and it has to cover at least 45% of the target sequence. The discrepancy in the alignment coverages reflects the possible inaccuracy in the start position of an ORF, where the boundaries of the previously annotated gene are considered to be more accurate.

A signal peptide prediction is considered to be reliable if its probability as reported by SignalP is  $\geq 0.75$  and its cleavage site probability is  $> 0.5$ .

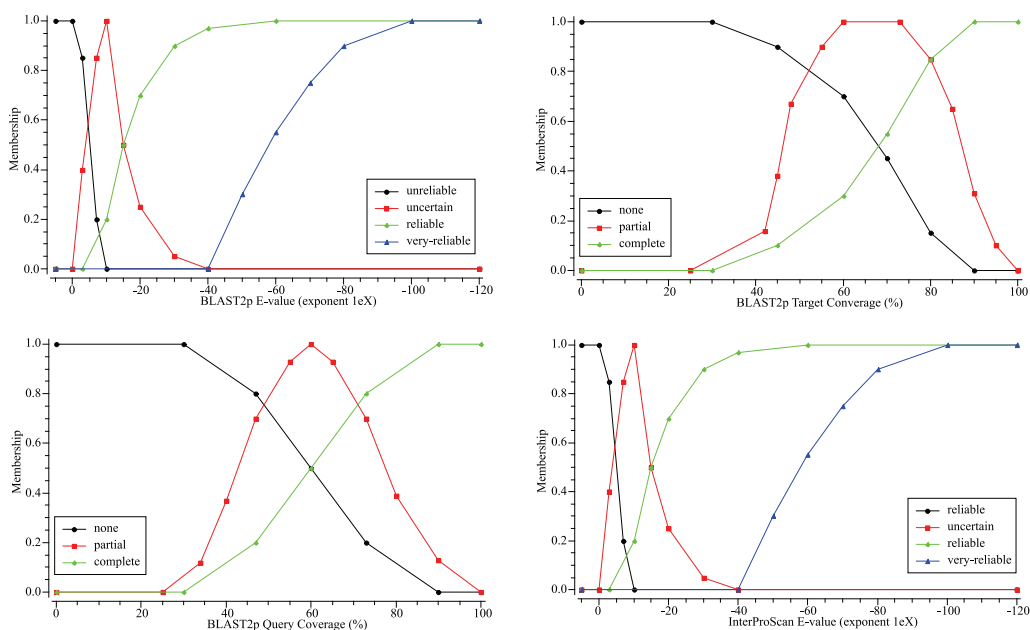
Transmembrane helix predictions reported by TMHMM are always regarded as reliable. Other observations reported by TMHMM are not considered later on.

**very\_reliable:** The class *very\_reliable* is a subclass of class *reliable*. As such its prerequisites have to be met and the E-value has to be  $< 1e^{-50}$ .

E-values closer to zero and coverages closer to 100% increase the reliability of an observation. The membership to the set *very\_reliable* of the linguistic variable *E-value* reaches its maximum of 1.0 at  $1e^{-100}$ . Smaller E-values do not contribute to the reliability of an observation any further. The membership functions for the linguistic variables used by MicHanThi are depicted in Figure 3.5

**Annotation** Annotations are created separately for the results of each class of tools. Observations reported by BLAST need the most processing. To derive a function based on these observations, the observations are grouped based on their functional description. Groups are ignored if the number of observations is too low when compared to the maximum number of observations in a group. Also, groups are ignored if the reliability of the supporting observations is too low when compared to the group supported by the most reliable observations. If a group is a subset of another group - all observations of group A are also included in group B - then the two groups are merged and the containing group is selected. Detailed information about this process can be found in Section 7.3.3.

Annotations created based on observations reported by InterProScan do not require that much effort. The prototype developed in (10) simply selected the



**Figure 3.5:** Membership functions for the different linguistic variables used in the for the evaluation of BLAST and InterPro observations.

most reliable observation that described a Pfam (52) protein family and created an annotation based on this observation. The enhanced version of MicHanThi now uses observations based on Pfam domains as well as observations based on the TIGRFAMs database, as well. Additionally, the new version considers the relation between two or more InterPro observations in the annotation process.

These relations are exploited by MicHanThi to select the most specific observation for the annotation of a gene. Parent entries are deleted if the child is at least 80% as reliable as the parent is. InterPro entries found in other entries are also deleted if the container is at least 80% as reliable as the contained entry. For each remaining InterPro observation, an annotation is created.

**Annotation Evaluation** Annotations based on the observations reported by different search tools are compared once all annotations are created. If two or more annotations describe the same function they are merged and missing information is transferred. Annotations are deleted if their reliability is too low compared to the most reliable annotation (80% of the most reliable annotation). For more information on how annotations based on observations created by different tools are merged see Section 7.3.3. The ORF is annotated as *hypothetical protein* if no functional annotation could be created.

**[Conserved] Hypothetical Proteins** If neither observations reported by BLAST nor observations reported by InterProScan are suitable for the annotation of an ORF then the ORF is annotated as *hypothetical protein*. If *uncertain* or more reliable observations were reported but these observations do not describe a gene function then the ORF is annotated as *conserved hypothetical protein*. In these cases the results obtained by the tools SignalP and TMHMM are considered

to assign the attributes *secreted* or *membrane*. The following list describes the rules used to assign these attributes.

**hypothetical protein:** An ORF is described as a *hypothetical protein* if no matches could be found in any of the sequence databases or if the reported matches are considered *unreliable*.

**conserved hypothetical protein:** The attribute *conserved* is assigned to a hypothetical ORF if at least one *uncertain* or better observation was found.

**protein containing:** If no reliable BLAST observations were found but a reliable observation describing an Pfam domain has been found, then an ORF is annotated as a *protein containing*.

**transmembrane prediction:** For ORFs that have at least two reliable transmembrane helix predictions the attribute *membrane* is assigned.

**signal peptide prediction:** If no more than one transmembrane helix was predicted and a reliable signal peptide prediction exists, then the ORF is annotated as *secreted*.

**transmembrane and signal peptide predictions:** If exactly one transmembrane helix prediction exists for an ORF and the predicted signal peptide prediction is uncertain because its HMM cleavage site probability is  $\leq 0.5$ , then the ORF is annotated as *membrane or secreted*.

### 3.2.2 Results

To validate the annotations proposed by MicHanThi the manually annotated genome of *Gramella forsetii* KT0803 (61) was used. Details about the performance of MicHanThi as compared to two automatic approaches for the functional annotation of genes can be found in Section 7.5.3.

The genome of *G. forsetii* - including 3593 ORFs - was chosen because extensive effort went into the manual annotation and its curation. It is believed that most ORFs in this genome are correctly annotated by the human annotators. As such, the annotation serves as *gold standard* for the evaluation of the developed tool. Another important factor is that the observations the manual annotations are based on were available and they could be accessed by the automatic annotation tools. Otherwise, annotations created by two different approaches could not have been directly compared as the information the annotations are based on differs.

To evaluate the performance of two approaches a script was written which compares the annotations created by each approach. This script compares the terms used in two annotations independent of their order. If all terms of annotation A are also used in annotation B then these two annotations are an exact match. The script also reports the number of subset matches. These are matches where one annotation uses a subset of terms used by a second annotation. In most cases the reason for a subset match can be found in the specificity of the annotations. The annotations *Glycosyl transferase* and *Glycosyl transferase, family N* are a typical example of a subset match where the second annotation is more specific than the first.

More information about the script used to evaluate the different tools can be found in Section 7.3.5. A summary of the results will be given in the following paragraphs. Please refer to Section 7.5 for a more detailed discussion.

**Evaluation of MicHanThi** Compared to the annotations created by human annotators, the annotations for 62% of the ORFs, predicted in the genome of *G. forsetii*, matched those created by the prototype of MicHanThi (10). Since then, MicHanThi was used in most annotation projects coordinated by the Microbial Genomics Group at the Max Planck Institute for Marine Microbiology in Bremen. Among these projects are the published genome studies of *O. algarvensis*, *Congregibacter litoralis* KT71, and four magneto tactic organisms (27; 73; 74), as well as the studies (75; 76).

Since the first successful application of MicHanThi in a genome annotation project the performance of the software has continuously been improved. The first improvement was the extension of MicHanThi to also include Pfam domain predictions as well as observations based on InterPro / TIGRFAMS. Another important contribution was the feedback gathered during the manual annotation of the organism *Congregibacter litoralis* KT71. This feedback was mostly used to improve the rule base as well as to adjust the thresholds used during the annotation process. After the improvements were implemented the annotations created by MicHanThi were again compared to the annotations manually created for the ORFs predicted in the genome of *G. forsetii*. Including the improvements, MicHanThi now creates annotations for 72% of the ORFs that match those created by the human experts.

During the implementation of the new version of MicHanThi it became apparent that slight differences in wording - such as *Glycoside hydrolase, family 17* and *Glycosyl hydrolases family 17* as used in InterPro and Pfam observations - caused the semi-automatic evaluation to wrongly consider annotations based on these observations as not describing the same function. To quantify the bias introduced by the spelling of words and by the use of synonyms, the first one hundred ORFs of the *G. forsetii* genome were manually inspected. This manual inspection revealed that roughly 8% of the created annotations have differences in wordings, but still describe the same function. MicHanThi created annotations for approximately 80% of the ORFs that can be compared to the annotations created by a human annotator considering these annotations

### 3.2.3 MicHanThi Accuracy / Human Inaccuracy

**Eye on Details** One of the biggest problems in the annotation of organisms is the level of experience of the annotators. The more experienced an annotator is, the more details will be considered for the annotation of an ORF. In the genome of *G. forsetii*, 1598 ORFs were initially assigned no function. After the annotations were corrected by human experts, the annotations of more than six hundred ORFs without a functional assignment had been changed. In most cases, the annotators neglected additional information such as predicted signal peptides and transmembrane helix prediction. In this class, the number of matching annotations per ORF increased by 44% from 826 to 1186.

**InterPro Relations** A new reason for mismatches in the annotation for the same ORF was introduced by the consideration of InterPro relations. In addition to an observation's reliability, the enhanced version of MicHanThi uses these relations to create the most specific annotation possible. An observation describing a broader function or a domain contained in another reported observation will not be used in the annotation process if the more specific observation is at least 80% as reliable as the less specific observation. Compared to the software, the human annotator commonly favoured the more reliable observation even if this observation is describing a broader function or a domain.

One reason for this behaviour can be found in the GenDB annotation system and InterPro itself. The annotation system presents the observations ranked according to the tool result. In this case the E-value reported by InterProScan. In cases where the less specific observation or an observation describing a single domain show a better E-value than more specific observations, then the less specific observations will be shown before the specific observations. Also, the information about the relation between two InterPro observations are not reported in the GenDB interface. To investigate the relation, an annotator would have to check the original InterPro entry at the InterPro website. From the information found at that source the annotators have to create a graph on their own. Based on this graph, they then have to decide if the most specific observation is, compared to all other observation, trust worthy enough to create an annotation based on it. For these reasons the more inexperienced annotators commonly based their annotation decisions on the most reliable observation instead of the most specific annotation.

## Chapter 4

# Of Avalanches and Tsunamis

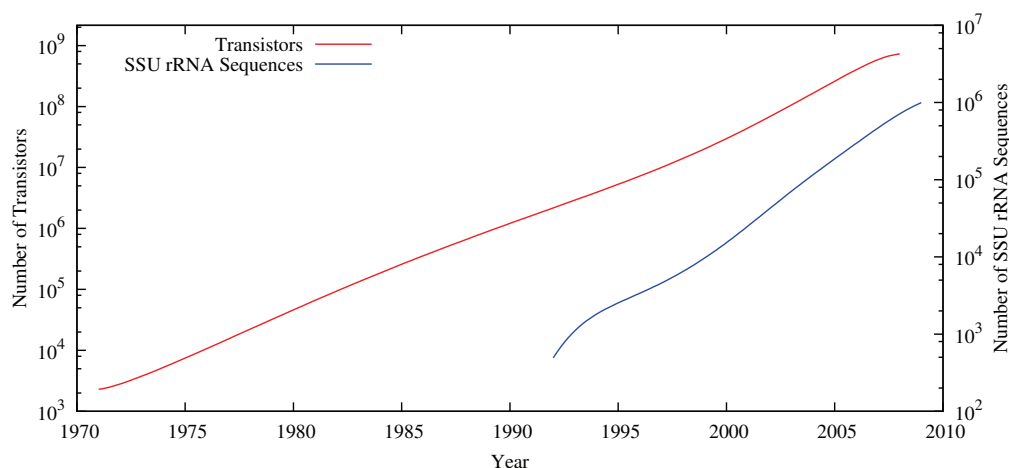
One of the most daring problems Bioinformatics faces today, is the exponential increase of publicly available sequence data since the early 1990s. Today, new sequencing robots are on the market, which rigorously break with the old sequencing concepts used by Sanger sequencing (23; 25; 26). These sequencers produce more genomic sequence data at a lower price in a shorter time. An end can, therefore, not be anticipated for the growth of publicly available sequence databases. Even more dramatic, the *rate* at which the available data doubles increases.

The first revolution in sequencing was initiated in the early 1990s by the human genome project which led to the assembly of enormous sequencing capacities. Large sequencing factories were built whose only purpose is the sequencing of biological samples. The sequencers in these factories use enhanced versions of the method developed almost two decades earlier by Frederick Sanger (13).

The second revolution was started by the development of pyrosequencing (2001) and particularly when 454 Life Sciences licensed this technology. In its first generation, this method produced a manifold of sequence data that could be obtained by classic sequencing methods. Since 454 Life Sciences licensed pyrosequencing in 2004, three generations of sequencing robots have been developed in less than five years. Each generation further decreasing the cost of sequencing and dramatically increasing the throughput as well as increasing the average read length. Two competing platforms are currently on the market, which compared to 454-pyrosequencing produce even more data, and new techniques are currently being developed or are already working in the laboratory, e.g. *Real-Time DNA Sequencing from Single Polymerase Molecules* (77). Two factors drive the rapid development of sequencing technology during this second revolution: the hunt for the *thousand-dollar human genome* (78; 79) and the ten million dollar denoted *Archon X PRIZE for Genomics* (80).

Before next generation sequencers entered the market, the increasing computer performance could be used to answer more and more complex questions. Now, the available and ever increasing computer performance is on the verge of being flooded by masses of sequence data that is increasing at an even faster rate (Figure 4.1). This greatly influences the requirements of tool development in Bioinformatics.

In the beginning, tool development was dominated by a brute force approach, adding more CPU's, main memory, and storage space. It did not even matter



**Figure 4.1:** Number of transistors used in Intel Desktop CPUs and the growth of SSU rRNA sequence databases (RDP 1992 – 2006 and SILVA SSU Parc 2007 – )

Sources:

[http://en.wikipedia.org/wiki/Transistor\\_count](http://en.wikipedia.org/wiki/Transistor_count),

[http://rdp8.cme.msu.edu/docs/rdp\\_release.html](http://rdp8.cme.msu.edu/docs/rdp_release.html),

<http://www.arb-silva.de/documentation/background/release-100/>

that tools were running for a few hours or days. Today, on the large data sets, the same tools would require weeks or even month to solve the problem. These initial tools were soon followed by more advanced tools that apply heuristics to reduce the computational complexity of a certain problem. In the following, some examples are given that demonstrate how the exponential growth of publicly available sequence data influences the development of tools.

## 4.1 Homology Searches

One example that describes the evolution of bioinformatics tools, from using the brute force approach to tools that use heuristics, can be found in the field of homology searches. The Smith-Waterman algorithm that is used to search for homologous sequences to a query sequence in a database. This algorithm creates ‘full’ local alignments for each sequence in the database and the query sequence. It is guaranteed to find the optimal local alignment between two sequences and it will find the most closely related sequence to the query sequence. In the early 1990s, searches, using the the Smith-Waterman algorithm, became computationally to expensive and tools were needed that could be used to search in the growing databases.

Out of this necessity the BLAST algorithm was developed (31). It is based on the same concepts as the Smith-Waterman algorithm but differs in two aspects: it uses a heuristic to reduce the search space, and it reports more results than the optimal alignment. However, the implementation of BLAST, e.g. shipped with Ubuntu 8.04 long term support (LTS) (version 2.2.17) faces its limitations, in form of *segmentation faults*, when it is used to search against databases which are larger than 4 GiB (4,294,967,296 bytes). The segmentation fault is caused



both on 32 bit CPU architectures as well as on 64 bit CPUs.

Two approaches can be used to solve this problem: databases larger than 4 GiB can be split into several smaller pieces, which are searched sequentially, and parallel computing methods can be used such as *massive parallel instruction (MPI)*. MPI allows to partition a search space, in this case the database, and different processes are used to search in these partitions in parallel (*parallel computing*). These processes may run on different computers on the same network or on the same multi CPU computer. The MPI environment orchestrates the different processes and merges the results. A drawback of using MPI is that it requires programs to be written specially for it, using custom MPI libraries. Also, not every problem can be partitioned because not all problems have a local memory footprint. A version of BLAST adapted for MPI is available from <http://www.mpiblast.org> (81).

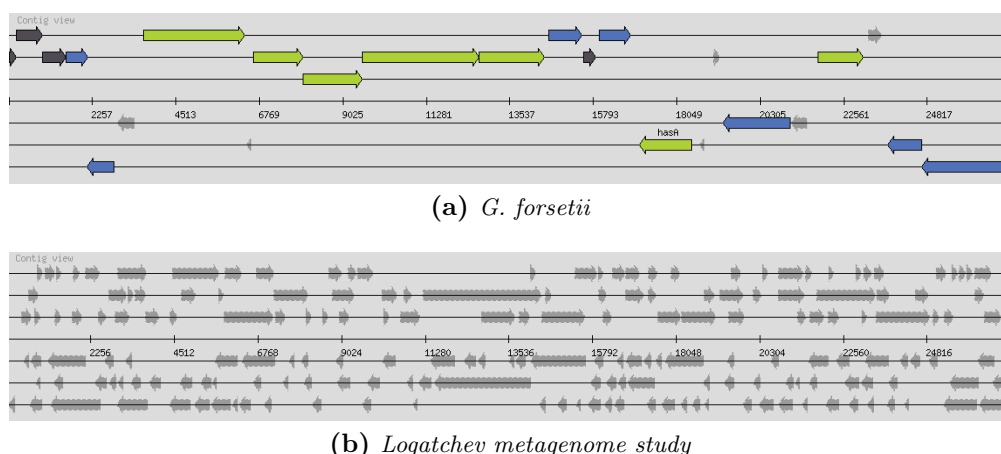
## 4.2 ORF Prediction

In 2003, Glöckner et. al annotated the organism *R. baltica* (21). One of the major challenges in this project was the ORF prediction, which lead to an over prediction of 45% (ORFs not submitted). As a consequence, they have started to develop a meta ORF finder which uses the results of several ORF finders. It applies a sophisticated reasoning to solve contradictions reported by different ORF finders and otherwise accepts ORFs predicted by all tools. For each ORF a plausibility check is made on top of the checks that are made by the independent ORF finders. Unfortunately, this tool never left the prototype state and could not be applied by non technical personal.

To give users the change to rely on more evidence than that reported by a single ORF prediction tool Jost Waldmann (Microbial Genomics Group – Max Planck Institute Bremen) developed the *porfs* script. This script takes a FASTA file as input and runs multiple ORF finders. Compared to MORFind it reports the results of all tools independently and it is left to the users to inspect the different ORF predictions and select those which are most reliable.

In currently studied metagenomes, e.g. the 454-pyrosequencing study of an environmental sample from the Logatchev sampling site (*Regina Schauer unpublished* – Department of Molecular Ecology, Max Planck Institute Bremen), the ORF prediction is done brute force. Brute force in this case means that the longest combination of start stop codon in each reading frame is accepted as possible ORF. Very short, very long ORFs, as well as overlaps between ORFs in different reading frames are not evaluated and need to be inspected manually. A typical ORF prediction for this project (Fig. 4.2b) and an exert of the ORF prediction in the *G. forsetii* (Fig. 4.2a) genome is shown in Figure 4.2.

Commonly there is an ORF predicted for every one thousand nucleotides in bacterial genomes. In the organism *R. baltica* more than thirteen thousand ORFs were predicted in a 7.1 MB genome. Of these ORFs, about 7,300 were finally submitted, 55% of the predicted ORFs. The following genome annotation projects used the MORFind (*Jost Waldmann unpublished*) tool to improve the ORF prediction. In these projects, all predicted ORFs were also submitted to the INSDC databases. In the Logatchev study, 113,299 ORFs have been predicted on 32,221 contigs which comprise 28 MB. The majority of ORFs (95,364 – 84%)



**Figure 4.2:** Two ORF predictions (viewed in GenDB): a) the closed genome of *G. forsetii* (Shotgun Sequencing), nucleotide positions: 1 – 27,079, and b) the Logatchev metagenome study (454-pyrosequencing), contig: contig00408, nucleotide positions: 1 – 27,079.

is less than 500 nucleotides long. 91,217 ORFs have no homologies in publicly available databases, and 25,494 ORFs could be automatically annotated using MicHanThi; 4,661 based on Swiss-Prot (47) observations, and 9,028 supported by InterPro (11) hits.

Besides the fact that the over prediction of ORFs makes it tedious to manually annotate the ORFs, it also has negative effects on the automatic annotation process, especially vertical annotation. Further problems caused by the over prediction of ORFs are the increased computer resources that are required to compute homologies. The bioinformatics compute cluster of the MPI Bremen includes eighteen 32 bit dual-Xeon nodes, five dual Opteron dual core nodes, and sixteen dual Opteron quad core nodes. Overall, more than two hundred compute slots are available. An average sized bacterial genome (4 MB) can be processed in roughly four hours (1,000 ORFs / hour), running the default set of tools that include BLAST against various databases, Pfam, and InterPro searches. Additionally, tools to search for signal peptides, transmembrane regions, tRNAs, and rRNAs are run. The processing of all ORFs in the Logatchev study took nearly five days and upcoming studies, in e.g. BMBF<sup>1</sup> project for the study of *Microrbial Interactions in Marine Systems (MIMAS)*, will comprise hundreds of mega bases. These projects will reach the limits of the available computer resources.

### 4.3 Representative Sets

The SILVA SSU Parc database, which now contains approximately one million SSU rRNA sequences, is used as a third example why large datasets pose a problem for day-to-day work in Bioinformatics. The size of the database causes three problems: the calculation of a phylogenetic tree, probe design, and the usage of the database on office PCs.

Building phylogenetic trees for almost one million sequences is currently computationally not possible. Even the number of sequences in the SILVA SSU Ref database is too large for the de novo calculation of a phylogenetic tree. The guide tree, included in the ARB export of this database, is created by incrementally adding small sets of sequences, newly added to the current release of the database, to the guide tree contained in the previous database.

The second problem is the design of probes for the identification and quantification of organisms in environmental samples. This requires large databases to assure a high sensitivity and a high specificity of the designed probes. A probe designed based on small databases or on databases missing complete groups of organisms may also detect members outside of the target group. This is caused by the fact that missing members of a group or missing groups of organism (even distantly related groups) may share certain regions of the DNA that, in their absence, is unique to the target group. For the design of probes a database must therefore either be as comprehensive as possible or the (selected) subset of sequences must be representative.

The obvious solution of the size problem is to only work with subsets of the sequences contained in the SSU databases. One approach to reduce the size of the database is to delete all sequences that are, according to one of the provided taxonomies, only distantly related to the groups of organisms a user is interested in. This leads to the problem related to probe design described above. A second approach is to reduce the size of the database by selecting a set of representative sequences for every organism group. Aside the problem that is caused by how to define the representative sequence for a given set of sequences, the computational complexity and space requirements for the involved clustering methods are enormous.

In clustering, a distance measure has to be established to define the relatedness between the objects to be clustered. In Bioinformatics, the alignment distance is one such measure. To apply the clustering algorithm, a matrix of the distances between all sequences must be created. This can either be done by creating a single multiple sequence alignment of all sequences or by creating pairwise sequence alignments of all sequences against all other sequences. Creating a MSA is only feasible for a few hundred or a couple of thousand sequences, otherwise, it is computationally too expensive.

The second approach requires a large number of pairwise sequence alignments which could be possible, depending on the available computing resources. However, the size of the resulting distance matrix would be  $n^2$ , where  $n$  is the number of sequences in the database. For hundreds of thousands of sequences, like the SILVA databases, the size of this matrix exceeds the limits currently provided by computer technology. In case of the SILVA SSU Ref database,  $409,907^2 * 8 \text{ bytes} = 1,344,189,989,192 \text{ bytes}$  ( $\approx 1.2 \text{ TiB}$ ) memory would be required to hold the complete matrix. This size could be reduced to  $\frac{n^2-n}{2}$  if the clustering algorithm uses a triangle matrix as input which assumes a symmetric distance between two sequences. The resulting 600 GiB of needed memory for the SILVA SSU Ref database still exceed the available resources in almost any case. A different approach needs to be found to select a representative set of sequences from SILVA databases.

One approach to reduce the available sequence set, by selecting representative

sequences, could be the approach followed by the ESPRIT tool (82). The authors assume that a very small set of 1% to 5% of the sequences is sufficient to estimate the biodiversity in biological samples. Instead of using the alignment distance it uses the k-mer distance of two sequences, which is less computationally expensive, to preselect a set of sequences which are then aligned and clustered. It creates a sparse distant matrix of all sequences and only keeps those results that meet a predefined threshold. Then it clusters the sequences based on their similarity score and last, it selects the representative sequence for each cluster. The tool can be distributed on a computer cluster to analyse a large number of sequences. While the exact algorithm followed by the tool may not be suitable to create a representative subset of sequences in the SILVA databases it may serve as a starting point for the development of such an algorithm.

## 4.4 Conclusions

Up to here, I drew a picture of the challenges Bioinformatics already faced in the past years and the challenges it faces in the upcoming years. One challenge or better said limiting aspect was completely neglected, so far: the human resource!

The problems Bioinformatics faces become more and more complex both on a computational level and on the evaluation side. The increasing databases cause an increasing complexity in the design of programs and in the compute infrastructure. Both, program development and infrastructure, require a lot of manpower which need to be considered during the planning stages of projects. Mayer et. al recently published a paper reviewing the usage and total cost of ownership of local compute clusters as compared to *cloud computing* (83). They estimate an average maintenance cost of  $\approx 1,400$  US Dollar per year for a single node in a compute cluster that consists of 128 nodes, excluding the cost for resources like electricity and cooling.

During the early design stage of any program, parallel and distributed computing techniques must be considered and the problem must be evaluated accordingly to check if any of these techniques can be applied to solve the problem. Normally, smaller tools that build on top of each other are more adequate to solve a problem and they are more flexible to use than monolithic tools that require user interactions. While these programs are written quite fast a lot of effort is needed to manage the dependencies between the programs and to distribute the programs on a compute cluster. In the SILVA project more time was spent to connect the different tools and to find a way to robustly run the complete SILVA pipeline as was spent to write the tools themselves.

During the implementation of a program the programming language, and the data storage are two important factors. The programming language must be able to efficiently work with large data sets that commonly require several GiB of main memory and even more persistent storage. Programmers need to possess detailed knowledge about the capabilities of the programming language and they have to make use of these possibilities to design and implement data structures efficiently. In most cases, only a fraction of the data stored on disk will fit into a computer's main memory and the program must be able to handle these cases, e.g. data structures / containers must support lazy loading.

All results that are produced by any of the bioinformatics tools need to be

inspected by biologists later on. The larger the results sets are the more complicated it is to ‘make sense’ of the data and the ‘traditional’ approach to inspect the data using graphical / web user interfaces, such as GenDB (9), and to keep track of the data in spreadsheets, which are limited in size, becomes more and more inadequate.

Graphical user interfaces that present all aspects of the data to a user are difficult to design. Biologists would need to turn to the authors of a program for all changes to the data views they require. A different user or a different project may require different views on the same data so that for each problem changes to the interface would have to be made. In our metagenome projects, the data are stored in relational databases and can be accessed using SQL. For both sides, it is much more beneficial to teach biologists basic computer skills like scripting and the use of SQL. The use of SQL allows biologists to easily answer their own questions without the need to ask bioinformaticians for the required data or to change the interface of a program, which would always cause delays. Bash scripting or simple Perl / Python scripting could be taught to allow biologists to automatise simple reoccurring tasks. This would further reduce the entanglement between the involved groups.

As a conclusion it could be said that bioinformaticians need to solve more and more complex problems using the limited available resources efficiently and that biologists need to learn basic computer skills such as the work on the command line, scripting, and SQL.



## Chapter 5

# Acknowledgments

Foremost, I would like to thank Prof. Dr. Frank Oliver Glöckner for giving me the opportunity to write this thesis and helping me through all the rough patches over the past three years. I would also like to thank Prof. Dr. Otthein Herzog for accepting this thesis within his group and for being the second supervisor of this thesis.

Many thanks to Elmar Prüße and the rest of the SILVA team without whom this thesis would not have been possible. A lot of effort in proofreading, was spent by Dr. Jörg Peplies, especially when it came to the paragraphs about phylogenetic analysis, many thanks for this. One of the largest problems during this thesis was not inalienable related to this work: the maintenance of the compute cluster. Many thanks to the IT department at the MPI and especially Carsten John for helping me maintaining the computers. I would also like to thank all the members of the Microbial Genomics Group and the Department of Molecular Ecology, not particularly mentioned here for just being around. Words of gratitude must be spoken to the L<sup>A</sup>T<sub>E</sub>X community for providing this great tool and the outstanding documentation. Even though it is hard at times, the results are always worth the costs.

And finally, many thanks to my family for the continued support of the past years.





**Part II**

**Publications**



# Table of Contents

---

<b>6</b>	<b>Silva Paper</b>	<b>55</b>
6.1	Introduction . . . . .	57
6.2	Materials and Methods . . . . .	58
6.3	Results and Discussion . . . . .	65
6.4	Conclusions . . . . .	69
6.5	Acknowledgments . . . . .	69
<b>7</b>	<b>MicHanThi Manuscript</b>	<b>71</b>
7.1	Introduction . . . . .	73
7.2	Methods . . . . .	74
7.3	Algorithm . . . . .	76
7.4	Design and Implementation . . . . .	82
7.5	Results . . . . .	83
7.6	Discussion . . . . .	87
7.7	Conclusions . . . . .	89
<b>8</b>	<b><i>Gramella forsetii</i> KT0803 Paper</b>	<b>91</b>
8.1	Introduction . . . . .	93
8.2	Results and Discussion . . . . .	94
8.3	Conclusions . . . . .	102
8.4	Experimental procedures . . . . .	103
8.5	Acknowledgements . . . . .	103
<b>9</b>	<b><i>Congregibacter litoralis</i> KT71 Paper</b>	<b>105</b>
9.1	Introduction . . . . .	107
9.2	Results and Discussion . . . . .	109
9.3	Materials and Methods . . . . .	115
9.4	Acknowledgments . . . . .	116

<b>10</b>	<b>Pirellula Paper</b>	<b>119</b>
10.1	Background . . . . .	121
10.2	Results and Discussion . . . . .	122
10.3	Conclusion . . . . .	129
10.4	Methods . . . . .	130
10.5	List of abbreviations . . . . .	133
10.6	Competing interests . . . . .	133
10.7	Authors contribution . . . . .	133
10.8	Acknowledgements . . . . .	133
<b>11</b>	<b>Megx.net Paper</b>	<b>141</b>
11.1	Introduction . . . . .	143
11.2	Sources of Genomic and Metagenomic Data . . . . .	144
11.3	Genome Browsing . . . . .	144
11.4	Precomputed Information . . . . .	144
11.5	TETRA Server . . . . .	145
11.6	Genomes Mapservers . . . . .	145
11.7	Additional Features . . . . .	146
11.8	Databases Access . . . . .	146
11.9	Acknowledgements . . . . .	146

---

## Chapter 6

# Silva Paper

### **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.**

E. Prbe<sup>a,b</sup>, C. Quast<sup>a,c</sup>, K. Knittel<sup>d</sup>, B.M. Fuchs<sup>d</sup>,  
W. Ludwig<sup>e</sup>, J. Peplies<sup>f</sup>, and F.O. Glckner<sup>a,c</sup>

<sup>a</sup>Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>b</sup>Center for Computing Technologies, University of Bremen, D-28359 Bremen, Germany;

<sup>c</sup>Jacobs University Bremen gGmbH, D-28759 Bremen, Germany;

<sup>d</sup>Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>e</sup>Department of Microbiology, Technical University Munich, D-85354 Freising, Germany;

<sup>f</sup>Ribcon GmbH, D-28359 Bremen, Germany;

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

**Journal; Volume (Issue):** Nucl. Acids Res.; 35 (21)

**Pages:** 7188-7196

**Month / Year:** September 2007

**DOI:** 10.1093/nar/gkm864

#### **Contributions:**

Design and implementation focused on the pipeline, IO library, database model and its abstraction layer, as well as the representation of the taxonomy. Prepared releases SILVA 89 through SILVA 95. Modification of the Phoenix EMBL parser (<http://www.bioinformatics.org/phoenix/wiki/>) to support compressed files in the bzip2 / gzip compression format, and adaptation to changes in the EMBL format. Now also official co-maintainer of the Phoenix EMBL parser.

---

**Contents**

---

<b>6.1</b>	<b>Introduction</b>	<b>57</b>
<b>6.2</b>	<b>Materials and Methods</b>	<b>58</b>
6.2.1	Sequence data	58
6.2.2	Quality checks	58
6.2.3	Aligner	59
6.2.4	Anomaly check	60
6.2.5	Taxonomy	60
6.2.6	Nomenclature	63
6.2.7	SSU and LSU rRNA databases for ARB	63
6.2.8	Availability / Webpage	63
6.2.9	Operating systems and programming languages	65
<b>6.3</b>	<b>Results and Discussion</b>	<b>65</b>
6.3.1	Data retrieval and processing	65
6.3.2	Alignment and aligner	68
6.3.3	Future developments	68
<b>6.4</b>	<b>Conclusions</b>	<b>69</b>
<b>6.5</b>	<b>Acknowledgments</b>	<b>69</b>

---

## Abstract

Sequencing ribosomal RNA (rRNA) genes is currently the method of choice for phylogenetic reconstruction, nucleic acid based detection and quantification of microbial diversity. The ARB software suite with its corresponding rRNA datasets has been accepted by researchers worldwide as a standard tool for large scale rRNA analysis. However, the rapid increase of publicly available rRNA sequence data has recently hampered the maintenance of comprehensive and curated rRNA knowledge databases. A new system, SILVA (from Latin *silva*, forest), was implemented to provide a central comprehensive web resource for up to date, quality controlled databases of aligned rRNA sequences from the *Bacteria*, *Archaea* and *Eukarya* domains. All sequences are checked for anomalies, carry a rich set of sequence associated contextual information, have multiple taxonomic classifications, and the latest validly described nomenclature. Furthermore, two precompiled sequence datasets compatible with ARB are offered for download on the SILVA website: (i) the reference (Ref) datasets, comprising only high quality, nearly full length sequences suitable for in-depth phylogenetic analysis and probe design and (ii) the comprehensive Parc datasets with all publicly available rRNA sequences longer than 300 nucleotides suitable for biodiversity analyses. The latest publicly available database release 91 (August 2007) hosts 547 521 sequences split into 461 823 small subunit and 85 689 large subunit rRNAs.

## 6.1 Introduction

Initiated by the pioneering studies of Fox and Woese (84) 30 years ago and later on pursued by Pace, Olsen, Giovannoni, and Ward (85; 86; 87; 88), the ribosomal RNA (rRNA) molecule has been established as the ‘gold-standard’ for the investigation of the phylogeny and ecology of microorganisms (19; 89). Today the more than 500 000 publicly available small and large subunit (SSU and LSU) rRNA sequences ask for specialized quality controlled databases and appropriate software tools. In anticipation of this impending deluge of rRNA data, the development of the ARB software suite and the curation of its associated databases began more than 12 years ago (1). The software suite offers a graphical user interface and a wide variety of interacting software tools built around a common database. Furthermore, the ARB project provides structured, integrative knowledge databases for small and large subunit rRNAs. Based on regularly offered international workshops and the ARB mailing list it is currently estimated that the ARB software suite and its databases are employed worldwide by several thousand users from academia and industry. In addition to the ARB approach, there are currently three projects offering access to a set of curated ribosomal RNA sequence and alignment databases: the European rRNA databank at the University of Gent (<http://www.psb.ugent.be/rRNA/>) (2) the Ribosomal Database Project II (<http://rdp.cme.msu.edu/>) at Michigan State University in East Lansing, MI (4; 56), and the greengenes project (<http://greengenes.lbl.gov/>) maintained by the Lawrence Berkeley National Laboratory in Berkeley, CA (3). All four projects offer at least one 16S rRNA dataset, but vary in the amount of sequences, quality checks, alignments, and update procedures. However, the ARB project is the only platform that actively incorporates homologous small (SSU) as well as large (LSU) subunit sequences from all three domains of life, the *Bacteria*, *Archaea* (16S/23S) and *Eukarya* (18S/28S). All projects provide web-based software tools for the alignment and classification of sequences as well as probe match functionalities. Downloading of sequences is provided in various formats including the commonly used FASTA and GenBank file formats. Additionally, greengenes provides ARB compatible datasets, but only for nearly full length sequences (>1250 bases) of *Bacteria* and *Archaea*.

An increasing awareness and motivation to catalogue and protect the biodiversity on Earth using molecular techniques demands comprehensive knowledge databases spanning all three domains of life. Furthermore, a majority of the sequences available is derived from cultivation independent biodiversity surveys, which rely on rapid pattern-

or clone-based approaches that often generate partial rRNA sequences. To conserve this suboptimal information especially for diversity studies, state of the art databases need to retain partial sequences.

To compensate for the limited phylogenetic resolution of the SSU rRNA (90; 91) the two fold larger LSU rRNA should now also be included in the rRNA approach (19). Especially for Eukaryotes the highly variable regions in the LSU rRNA are already commonly used for species discrimination (92). Triggered by a new capacity for cheap and rapid sequencing, there is a steady flow of approximately 10 000 rRNA sequences per month into the public databases of the International Nucleotide Sequence Database Collaboration (<http://www.insdc.org>). To make full use of these data for reliable phylogenetic reconstructions and biodiversity analysis careful inspection of each sequence and alignment is necessary. To support the users with this task, standardized procedures to assign a defined set of contextual information to each sequence must be established. Unified quality control mechanisms are urgently needed to intuitively flag potential problems with each sequence. The recent introduction of accelerated and less expensive sequencing technologies, such as pyrosequencing (93), and their successful application for a census of marine microbial diversity (94) further substantiates the need for comprehensive quality controlled databases for comparisons. Such databases provide a stable framework enabling biologists to transfer the copious data into reliable biological knowledge. The SILVA database project is designed to satisfy the request for comprehensive quality controlled and aligned rRNA datasets. It is intended to provide a central knowledge resource to alleviate users of the time consuming manual curation process.

## 6.2 Materials and Methods

### 6.2.1 Sequence data

The SILVA release cycle and numbering corresponds to that of the EMBL database, a member of the International Nucleotide Sequence Database Collaboration (<http://www.insdc.org>). Thus, the ribosomal RNA sequences used to build version 91 of the SILVA databases, which is referred to in this paper, were retrieved from release 91 (June 2007) of EMBL. A complex combination of keywords including all permutations of 16S/18S, 23S/28S, SSU, LSU, ribosomal and RNA was used to retrieve a comprehensive subset of all available small and large subunit ribosomal RNA sequences. All candidate rRNA sequences extracted from the EMBL database were stored locally in a relational database system (MySQL). The specificity of the SILVA databases for rRNA is assured by the subsequent processing of the primary sequence information.

The source database providing the seed alignment, required for the incremental alignment process, included a representative set of 51 601 aligned rRNA sequences from *Bacteria*, *Archaea* and *Eukarya* with 46 000 alignment positions. The SSU alignment positions are currently kept identical with the `ssu_jan04.arb` database which has officially been released by the ARB project (<http://www.arb-home.de>) in 2004. For the large subunit RNA databases, an in-house, aligned database was used as the seed. It encompasses a representative set of 2868 sequences from all three domains (150 000 alignment positions). Since the quality of the final datasets critically depends on the quality of the seed alignments both datasets were iteratively cross-checked by expert curators during database build-up. Within this process, all sequences that could not be unambiguously aligned were removed from the seed.

### 6.2.2 Quality checks

Every imported SSU and LSU sequence had to pass a multi-stage quality inspection. Sequences were rejected if they were shorter than 300 unaligned nucleotides, if they were composed of more than 2% of ambiguities or more than 2% homopolymeric stretches longer than four bases, which means only bases exceeding homotetramers are counted,



or if they had more than 5% identity to vector sequences. The identity was checked by querying a database of commonly used vector sequences, based on the EMVEC (<http://www.ebi.ac.uk/blastall/vectors.html>) and UniVec (<http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html>) databases using the blastn tool. All thresholds to reject sequences were defined based on statistical analysis of the retrieved SSU and LSU sequences. Each sequence in the SILVA databases carries the percentages of ambiguities, homopolymers, and vector contamination. Additionally, a summary ‘sequence quality’ score is calculated according to the following formula, where Sq = sequence quality, A = % ambiguities, H = % homopolymers and V = % vector identity:

$$Sq = 1 - \frac{\frac{A}{A_{max}} + \frac{H}{H_{max}} + \frac{V}{V_{max}}}{3} * 100 \quad (6.1)$$

This score represents the mean of the three individual parameters, such that 100 is the best possible value. All sequences that passed the quality thresholds were automatically aligned against the seed alignment using the new SILVA INcremental Aligner (SINA).

### 6.2.3 Aligner

To cope with the huge amount of sequence information and to minimize the workload for manual curation, a new dynamic incremental profile sequence aligner (SINA) was developed. In the first step the aligner uses the suffix tree concept of ARB (1) to search for up to 40 closely related sequences in the seed alignment. The reference sequences from the seed are transferred into a partial order graph as used in (95), while preserving the positional identity from the reference alignment. The sequence under investigation is then aligned to this graph using a variant of the Needleman-Wunsch algorithm (38) with affine gap penalties and cost free overhang. The graph concept allows ‘jumping’ between the different references to find an optimal alignment for the different sequence regions. This technique enables the algorithm to correctly place bases that were e.g. deleted from the closest relative, but are present in the candidate sequence and other relatives. It also eliminates the need for synthetic full length sequences in the reference alignment as introduced for the NAST aligner (96) To further improve the alignment quality a variability statistic is used to give more weight to conserved positions. Results of each step of the aligner are reported to the database and shown in the corresponding fields of the exported ARB file (Tables 6.1 – 6.3). The ‘alignment quality’ score is a measure of the similarity with the seed sequences that are taken into account for the alignment process. The score is derived from the dynamic programming score by removing the effects of sequence length and positional weighting. High values (>90) mean that nearly identical

**Table 6.1:** *Description of database fields in ARB files exported from SILVA for ARB specific fields and entries.*

ARB Name	Field	Owned By	Description
aligned		User	User-defined entry, e.g. name and date of the person who aligned the sequence
ambig		ARB	Ambiguities calculated in ARB using ‘count ambiguities’
ARB_color		ARB	Stores the information about sequence colors
name		ARB	Internal ARB database ID, do not change!
nuc		ARB	Number of nucleotides; calculated by ARB using ‘count nucleotides’
nuc_term		ARB	Number of nucleotides coding for the respective rRNA gene; calculated by ‘count nucleotides gene’
remark		User	Field for remarks
tmp		ARB	Used by several ARB modules

sequences have been found within the seed alignment, resulting in a high likelihood for the alignment to be accurate. Low values indicate a high distance as perceived by the aligner, making the alignment task more difficult and lowering the average accuracy. Due to the size of the seed alignment, low values are rather rare and ask for manual inspection of the alignment. The ‘basepair’ score is calculated from the number of bases involved in helix binding according to the secondary structure model of Gutell et al. (97) as already implemented in the ARB package. Canonical and non-canonical base pairings are evaluated, weighted according to the cost model implemented in the Probe.Match (‘weighted mismatches’) tool in ARB (1). To fit our unified scoring scheme, the alignment quality and the base pair score were normalized to values between 0 and 100, such that 100 represents the maximum score. After aligning, the number of successfully aligned bases was again counted and sequences with less than 300 bases within the boundaries of the respective SSU or LSU rRNA genes were discarded.

#### 6.2.4 Anomaly check

To check for sequence anomalies, a custom version of the Pintail software (59) was used. The software was specifically adapted for batch processing by the RDP II team. By design, Pintail can only detect anomalies between two sequences. To circumvent this limitation, a pairwise comparison of all sequences in the seed against a group of 20 sequences was performed. If a majority of the comparisons was deemed anomalous, the sequences were iteratively eliminated from the seed alignment until no such sequences remained. Subsequently, all aligned sequences of the SSU database were tested against their five closest relatives within this pruned seed. The number of ‘yes’, ‘likely’ and ‘no’ reported by Pintail was counted for each sequence and transferred into the ‘Pintail quality’ value. This score was normalized between 0 and 100, such that 100 indicates the best quality and a low probability that the sequence is anomalous or chimeric. Only SSU sequences were checked for anomalies because the Pintail software is currently not designed to handle LSU sequences.

#### 6.2.5 Taxonomy

Every sequence in the SILVA databases carries the EMBL taxonomy assignment. Where available, the greengenes and RDP taxonomies were added for comparison. The EMBL taxonomy was retrieved simultaneously with the sequence, whereas the other taxonomies have been assigned to the sequences based on accession numbers. The greengenes taxonomic outline was acquired in June 2007 from the greengenes website (<http://greengenes.lbl.gov/>) and the RDP Nomenclatural Taxonomy was acquired from RDP II release 9.51. At the moment, no other up to date databases containing aligned LSU sequences are available. Therefore, the only taxonomy provided with the LSU database is the taxonomy used by EMBL. Type strain information has been added to the field ‘strain’ and is indicated by [T]. Mapping was done based on the RDP II dataset and is therefore only available for *Bacteria*.

**Table 6.2:** *Description of database fields in ARB files exported from SILVA for Fields and entries imported from EMBL.*

<b>ARB Field Name</b>	<b>EMBL Field</b>	<b>Description</b>
acc	AC	Accession number
ali_xx/data	sequence	Sequence information
author	RA	Reference author(s)
clone	FT/clone	Clone from which the sequence was obtained
collected_by	FT/collected_by	Name of the person who collected the specimen
collection_date	FT/collection_date	Date that the specimen was collected
country	FT/country	Geographical origin of sequenced sample
date	DT	Entry creation and update date separated by;
description	DE	Description
full_name	OS	Organism species
gene	FT/gene	Symbol of the gene corresponding to a sequence region
insdc	PR	The International Nucleotide Sequence Database Collaboration (INSDC) Project Identifier that has been assigned to the entry
isolate	FT/isolate	Individual isolate from which the sequence was obtained
isolation_source	FT/isolation_source	Describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived
journal	RL	Reference location
lat_lon	FT/lat_lon	Geographical coordinates of the location where the specimen was collected
nuc_region	FT source	Identifies the biological source of the specified span of the sequence
nuc_rp	RP	Reference positions
product	FT/product	Name of the product associated with the feature
publication_doi	RX	Cross-reference DOI number
pubmed_id	RX	Cross-reference Pubmed ID
specific_host	FT/specific_host	Natural host from which the sequence was obtained
specimen_voucher	FT/specimen_voucher	An identifier of the individual or collection of the source organism and the place where it is currently stored, usually an institution
start	FT rRNA	Start of the ribosomal RNA gene
stop	FT rRNA	Stop of the ribosomal RNA gene
strain	FT/strain	Strain from which the sequence was obtained
submit_author	RL	Submission authors from reference location

Continued on next page

Table 6.2 – continued from previous page

ARB Field Name	EMBL Field	Description
submit_date	RL	Submission date from reference location
tax_emb1	OC	Organism classification according to EMBL
tax_emb1_name	OC	Organism name taken from the classification field
tax_xref_emb1	FT/db_xref	Database cross-reference: pointer to related information in another database
title	RT	Reference title
version	ID SV	Subversion from identification line

### 6.2.6 Nomenclature

All organism names have been synchronized with the ‘Nomenclature up to date’ website of the “Deutsche Sammlung für Mikroorganismen und Zellkulturen” DSMZ (released June 2007, <http://www.dsmz.de/download/bactnom/names.txt>) in order to stay consistent with the constant renaming of validly described species according to the recommendations published in the ‘International Journal of Systematic and Evolutionary Microbiology’. All former names are stored in the database and are visible on the web page, as well as in the corresponding field of the ARB databases (Tables 6.1 – 6.3).

### 6.2.7 SSU and LSU rRNA databases for ARB

Two types of precompiled databases for both small and large subunit ribosomal RNA sequences are available in the ARB format: the high-quality Ref databases and the comprehensive Parc databases. The Ref databases are subsets of Parc, which are exclusively comprised of nearly full length 16S/18S and 23S/28S rRNA sequences. A sequence is accepted if it is at least 1200 bases long. Additionally, sequences as short as 900 bases are included if they belong to the domain *Archaea*. Applying a strict cut-off at 1200 bases would result in the loss of the majority of sequences of this domain. Sequences in the LSU Ref database have to be at least 1900 bases long. For quality control, all sequences that could not be unambiguously aligned (alignment quality score <50 and <30 for SSU and LSU, respectively) were removed from the Ref databases. Both Ref databases are supplemented with a guide tree based on the full length sequence tree of the ARB Jan 04 SSU and the Ludwig LSU databases, respectively. The trees were built using the ARB parsimony tool with filters to remove highly variable positions. Common filters like the positional variability filters were recalculated for the Ref databases. Sequences with long branches in combination with low alignment qualities (<80) were removed from the Ref databases.

The rRNA Parc databases are a collection of all quality checked and automatically aligned rRNA sequences longer than 300 bases of the aligned rRNA gene (field ‘nuc\_gene\_slv’, Tables 6.1 – 6.3). The name Parc has been chosen according to the UniProt concept (51) where Parc stands for the comprehensive protein sequence archive. All sequences in the SILVA databases are associated with a rich set of sequence and process parameters. Included is information from the initial quality checks to the alignment process, as well as information taken directly from the EMBL entry (Tables 6.1 – 6.3). Together with the search and query functionalities on the web site and in ARB, one can quickly search for problematic sequences or generate individual high or low quality sequence subsets for further processing or curation. The ARB package can be used to export sequences in various formats like EMBL, GenBank, or aligned and unaligned FASTA.

### 6.2.8 Availability / Webpage

The SILVA databases are available via a web-based interface at <http://www.arb-silva.de>. The web interface is divided into six sections: the browser, search, list, download, background, and FAQs pages. Download of the complete Parc and Ref datasets in ARB format is available in the download section. It is also possible to download files that gain additional sequences from new releases. Subsets of aligned sequences from the Parc dataset can be retrieved from the website. This is currently possible via two entry points: a taxonomic browser and advanced search functions. After selecting a database and the desired taxonomy in the browser, the user can navigate through the taxonomy by clicking on the respective nodes. A cart system is used to easily select subsets of single sequences, complete groups or even phyla. The cart system keeps the selections from the SSU and LSU databases distinct. This allows the user to select sequences from both databases simultaneously without mixing the two sequence types. However, it must be

**Table 6.3:** *Description of database fields in ARB files exported from SILVA for SILVA specific fields and entries.*

<b>ARB Field Name</b>	<b>Description</b>
align_bp_score_slv	Calculates the number of bases in helices in the aligned sequence taken into account canonical and non canonical basepairing. The cost matrix is taken from ARB Probe-Match (1).
align_cutoff_head_slv	Unaligned bases at the beginning of the sequence
align_cutoff_tail_slv	Unaligned bases at the end of the sequence
align_family_slv	Names and scores of reference sequences in the alignment process
align_log_slv	Detailed aligner comments
align_quality_slv	Maximal similarity to reference sequence in the seed
aligned_slv	Data and time of alignment by Silva
ambig_slv	Calculated percent ambiguities in the sequences, a maximum of 2% is allowed
homop_slv	Calculated percent repetitive bases with more than four bases, a maximum of 2% is allowed
homop_events_slv	Absolute number of repetitive elements with more than four bases
nuc_gene_slv	Aligned bases within gene boundaries
pintail_slv	Information about potential sequence anomalies detected by Pintail (59); 100 means no anomalies found.
alternative_name_slv	Synonyms or basonyms of the species according to the DSMZ 'nomenclature up to date' catalogue
seq_quality_slv	Summary sequence quality value calculated based on values from vector, ambiguities and homopolymers, 100 means very good
tax_gg	Taxonomy mapped from greengenes
tax_gg_name	Organism name in greengenes
tax_rdp	Nomenclatural taxonomy mapped from RDP II
tax_rdp_name	Organism name in RDP II
vector_slv	Percent vector contamination, a maximum of 5% is allowed

noted that any misclassification or erroneous information provided by INSDC is currently propagated on the SILVA webpage.

Additionally, the advanced search functions of the SILVA website can be used to build custom subsets of sequences. In addition to simple searches e.g. for accession numbers, organism names, taxonomic entities, or publication DOI/PubMed IDs, complex queries over several database fields using constraints such as sequence length or quality values are possible. The results can be sorted according to accession numbers, organism names, sequence length, sequence and alignment quality and Pintail values. Before download, the search results must be added to the 'List'. This can be done by either manually selecting the sequences by mouse click or by clicking on 'Add complete result to List' to mark and transfer all results.

The coloured bars on the search page and in the short and detailed sequence views of the browser given a fast overview of the different quality aspects assigned to every sequence. The length of the bars is a graphical representation of the respective quality value. The colours classify the information into four categories: A green bar represents a value equal to or greater than 75. Yellow bars stand for values equal to or greater than 50 but less than 75. Values less than 50 are expressed by an orange bar. Red bars are only used for scores of 0. Since 'problematic' sequences, sequences of inadequate quality, as well as insufficiently aligned sequences were discarded from the databases only the Pintail scores can have 0.

In the 'List' section of the website, the entries can be inspected, items can be deleted, and the download files can be created. By clicking on the 'generate download' button the user will be asked whether he would like to download the sequences as a multi-FASTA or ARB file from the download section of the web page. All generated files will be available for download on the download page for up to 24 h. The background section of the website provides additional information about the current status of the databases, and the FAQ section describes the main steps necessary to download subsets of sequences and how to merge the retrieved ARB databases with the user's personal ARB database.

### 6.2.9 Operating systems and programming languages

The SILVA core system was written in C++ and runs on an Ubuntu GNU/Linux 6.06 LTS based 64bit Dual Dual-Core Opteron cluster with at least 16 GB of main memory on each node. The database server runs MySQL 5.0 and features 32 GB of main memory. The Sun-grid engine (N1GE 6.0) is used to distribute jobs, such as importing, quality check, and aligning on the cluster. The web server is a LAMP system running Ubuntu GNU/Linux 6.06 LTS, Apache 2, MySQL 5.0, and PHP 5. It is connected to the internet via a synchronous 34 Mb connection. The website was written in PHP and Ajax and it is managed using the typo3 content management system in version 4.1. Due to the complexity of the system and the high hardware requirements the system is currently not intended for local installation.

## 6.3 Results and Discussion

### 6.3.1 Data retrieval and processing

The total numbers of retrieved sequences and the number of and reasons for rejected sequences are listed in Table 6.4. Cross checks with RDP II and greengenes indicated a sensitivity of our search procedure of >99%. For ambiguities, homopolymers and vector contamination the numbers are non-additive, since some of the sequences may be affected by two or three parameters. Cut-off values have been determined based on a statistical evaluation with relaxed parameters (data not shown), and are intended to balance the quality of the databases with any loss of information. Manual inspection of the sequences rejected by the aligner showed that most of these sequences were not ribosomal RNA sequences.

**Table 6.4:** *Sequence retrieval and processing for SILVA 91*

	SSU Parc	LSU Parc
Candidates	900 573	417 217
<300 Bases	320 327	297 218
>2% Ambiguities	8018	2193
>2% Homopolymers	19 240	4772
>5% Vector contamination	14 973	2573
Insufficient relatives	49 063	13 081
<300 Gene bases	25 961	7510
<30 Alignment quality or base pair score	6583	3390
Total sequences in Parcs	461 823	85 689

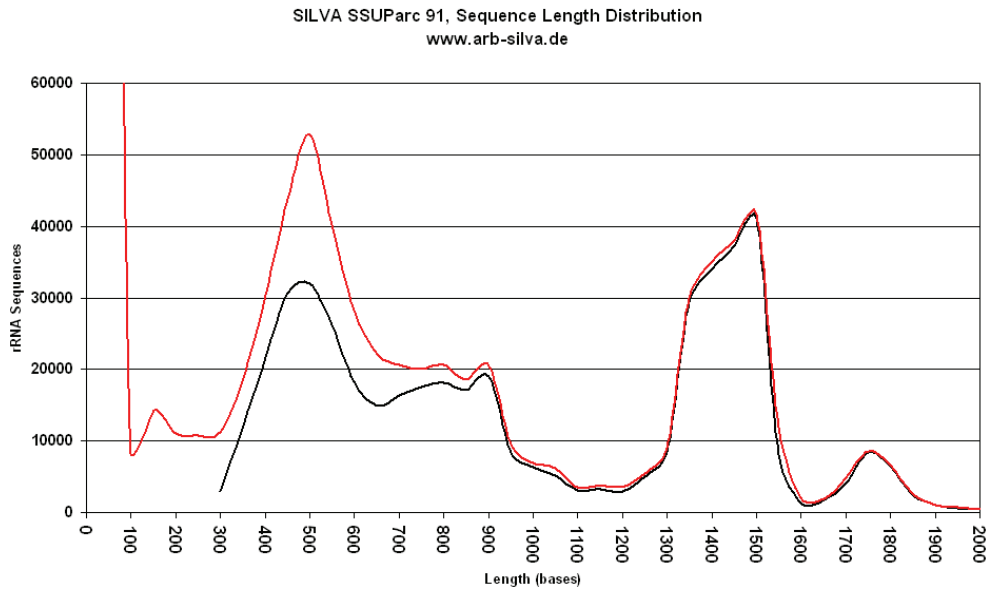
A comparison of the length distribution immediately after importing the SSU sequences with the length distribution of aligned sequences confirmed that no unexpected loss of sequences in certain length classes occurred (Figure 6.1). Partial sequences between 300 and 800 bases were more frequently rejected than longer ones. A closer comparison of sequence quality versus sequence length confirmed that sequences below 700 bases tend to be of low quality. These ‘problematic’ sequences may be generated in diversity studies based on single strand sequencing. The high number of rejected sequences with less than 300 bases is evidence for the increase in short length tag sequencing using e.g. pyrosequencing machines. The LSU database shows a similar distribution for rejected sequences as the SSU database (Figure 6.2).

As expected, the SSU sequence length distribution follows the prominent primer sets used for sequencing specific conserved regions on the 16S/18S rRNA gene (98). A distinct peak exists around 500 bases, a small one at 900 bases, and a peak between 1300 and 1500 bases. The large number of sequences with 300 and 600 bases is typical for diversity studies that use single reads or fingerprint techniques like DGGE (99). A text search for ‘DGGE’ across all fields of the SSU Parc database using ARB showed that 8241 (93%) out of 8889 ‘DGGE’ sequences found belong to the 300 – 600 nucleotide length class. A taxonomic breakdown for the 300 to 600, 600 to 1000, and 1300 to 1600 bases length classes revealed that 80 to 90% of all sequences per class were of bacterial origin. Nevertheless, from the shortest to the longest length class, the relative numbers for *Eukarya* decreases, whereas *Archaea* and *Bacteria* peaked in the 600 – 1000 and 1300 – 1600 length classes, respectively. This again reflects the application of the typical universal primers for *Bacteria* (98) and *Archaea* (100).

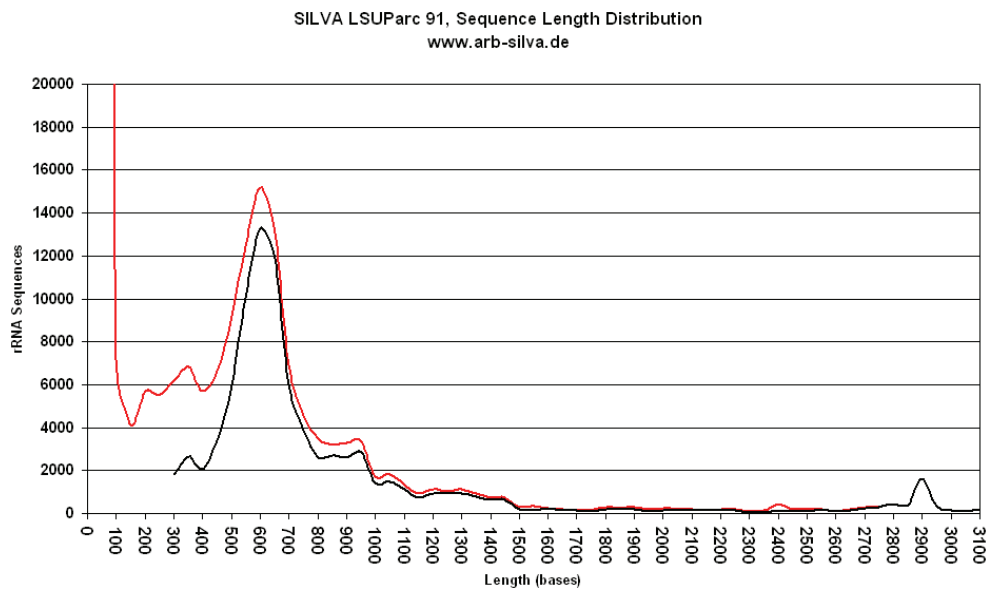
A comparison of the number of sequences hosted by the SILVA, greengenes, and RDP II projects revealed that the SILVA SSU Ref database contains roughly the same amount of bacterial and archaeal sequences as greengenes (3) [SILVA: 165 928, greengenes: 165 759 (July 2007)]. Furthermore, SILVA contains 2423 more nearly full length sequences for *Bacteria* than RDP II (163 505, release 9.52) (56). This is surprising considering SILVA’s less frequent release cycle (currently synchronized with major EMBL releases); one would thus anticipate SILVA to contain fewer sequences. This may have been due to a higher sensitivity in SILVA’s search and alignment protocol. Different quality control mechanisms should not have a significant influence, since only nearly full length sequences have been taken into account for this comparison.

With this respect it has to be emphasised that the primary intention of the SILVA project is not to offer the biggest database by size but more importantly to provide reliable rRNA datasets with a robust set of processing and quality values assigned to each sequence. Such quality values enable users to easily evaluate sequences in order to create subsets of sequences for specific applications, or to extract the sequences that need further attention with respect to sequence and/or alignment quality or anomalies. The alternative taxonomies and type strain information, as well as the latest nomenclature, will facilitate the daily work flow of diversity analysis using classical clone based and high





**Figure 6.1:** Sequence length distribution of rRNA genes in the SILVA 91 SSU database. The dotted line represents the sequence distribution directly after importing, the solid line after quality checks and alignment. The huge amount of sequences around 100 bases reflect the first impact of tag sequencing approaches.



**Figure 6.2:** Sequence length distribution in the SILVA 91 LSU database. The dotted line represents the sequence distribution directly after importing, the solid line after quality checks and alignment. The huge amount of sequences around 100 bases reflect the first impact of tag sequencing approaches.

throughput sequencing approaches. Additionally, SILVA provides two LSU databases to support the increasing use of molecular markers with a higher resolution than the SSU rRNA (90). A taxonomic breakdown of the LSU Parc database contents showed that already 91% of the sequences are of eukaryotic origin.

### 6.3.2 Alignment and aligner

The current SILVA alignment is based on 46 000 and 150 000 alignment positions for the small and large subunit rRNA, respectively. The reasons for the large amount of alignment positions are: (i) large insertions often present in *Eukarya* and (ii) sequencing errors, such as additional artificial bases often found in homopolymeric sequence stretches. Such errors are common and require placement to be filtered before phylogenetic tree reconstruction, without corrupting the rest of the alignment.

In the ‘align-to-seed’ approach implemented in the SILVA system, well aligned sequences from seed datasets are used as references for new, unaligned sequences. Therefore, the quality of the final alignment strongly depends on the accuracy of the seed alignment. To further improve the quality of the SSU and LSU seed databases a manual curation process was performed on the latest officially released ARB dataset from January 2004.

The SSU seed hosts currently over 1000 unpublished sequences that primarily cover the domain *Archaea*. These sequences further improve the alignment in regions of the original SSU January 2004 dataset with sparse sequence coverage. In summary, the quality and consistency of all of the seed alignments is excellent. Only minor inconsistencies could not be resolved in the *Eukarya*. Nevertheless, the Parc datasets exceed the corresponding SSU and LSU seeds by a factor of 8 to 25. This probably indicates that not every phylum is equally represented in the seed. Hence, before using the alignments for in-depth phylogenetic analysis, the alignment of the selected sequence should be carefully checked. Problematic sequences can be easily filtered out by the quality values followed by manual curation. The SILVA team highly appreciates the return of manually inspected and corrected alignments of sequence subsets for inclusion in the SILVA seed. This will allow us to further increase the quality of future alignments.

To manage the deluge of data currently available in the public databases, a new aligner (SINA) has been developed. Similar to existing aligners, such as the Fast Aligner implemented in ARB (1) or the NAST aligner (96), the tool uses related sequences from the reference alignment as a template. For benchmarking the performance of SINA, standard tools, such as BALiBASE (101), could not be used since they are restricted to protein sequences. Benchmark results were obtained by removing and realigning each sequence from the seed. The results were internally compared to the original alignment by counting the number of matching and non-matching columns. Overall, SINA correctly placed 99.8% of all bases in the alignment. Furthermore, 33% and 80% of all sequences tested had no, or less than 1%, alignment errors, respectively. The high accuracy was gained in extensive test runs by changing parameter sets for gap insertions/extension parameters and family sizes combined with subsequent manual inspection of the results by expert curators. The design and implementation of SINA as individually running processes allows distributed aligning on cluster nodes. More than one sequence per second can be aligned per CPU.

### 6.3.3 Future developments

To account for the growing awareness in ecology that sequence information must be treated in the proper environmental context (102), emphasis was put on the retrieval of contextual (meta)information from public databases. For easy visualisation, the ‘Environment’ subsection is available in the detailed view of the browser. Additionally, basic environmental parameters, such as exact location and time of sampling as well as physical, chemical, and biological information about the sampling site, will be added in

collaboration with the International Census of Marine Microbes (ICoMM), where similar efforts are ongoing (<http://icomm.mbl.edu/>). In upcoming releases of the SILVA databases a crosslink to the genomes mapserver at <http://www.megx.net> (66) will allow the geographic visualization of the sequence information as long as the location is provided. The direct addition of tag sequences below 300 nucleotides as typically produced by pyrosequencing, is not currently planned for SILVA, since it is already a main objective of the ICoMM agenda (94). Sequence based search options and alignment of user provided sequences are under development for the SILVA webpage. Finally, it must be stressed that an appropriate and stable phylogenetic classification of all rRNA sequences is urgently needed. Efforts in collaboration with Bergey's trust are ongoing and the information will be incorporated as soon as it becomes electronically available.

## 6.4 Conclusions

The new SILVA system provides comprehensive, quality controlled, richly annotated and aligned, reference rRNA databases to support the molecular assessment of biodiversity, as well as investigations of the evolution of organisms. Applications of the databases range from basic research in microbiology and molecular ecology to the detection of contaminants and pathogens in biotechnology and medicine. Molecular taxonomy and diagnostics have already revolutionized our view on microbial diversity on Earth (94; 103; 104), and the added value of molecular techniques for the determination of eukaryotic diversity has recently been documented by Tautz et al. (105) The SILVA databases combined with the ARB software suite provide a stable and easy to use workbench for researchers worldwide to perform in depth sequence analysis and phylogenetic reconstructions. It is designed as a knowledge database to assist in the daily effort to keep pace with the increasing amount of data flooding our general-purpose primary databases.

## 6.5 Acknowledgments

We would like to thank Ralf Westram for expert assistance with the ARB software suite, the company Pixelmotor for designing and implementing the webpage and all colleagues and students who helped with the manual curation of the databases. We would also thank James Cole, George Garrity and the RDP II team for help with Pintail and fruitful discussions. We are grateful for funding from the Max Planck Society. Funding to pay the Open Access publication charges for this article was provided by the Max Planck Society.



# Chapter 7

## MicHanThi Manuscript

### MicHanThi – Automated Prediction of Gene Functions.

C. Quast<sup>a,b</sup>, I. Kostadinov<sup>a,c</sup>, O. Herzog<sup>b</sup>,  
and F.O. Glöckner<sup>a,c</sup>

<sup>a</sup>Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>b</sup>University of Bremen, Center for Computing Technologies (TZI), D-28359 Bremen, Germany;

<sup>c</sup>Jacobs University Bremen gGmbH, D-28759 Bremen, Germany;

**Journal; Volume (Issue):** In Preparation

**Contributions:**

Concept, design, implementation, evaluation, and script.

---

**Contents**

---

<b>7.1</b>	<b>Introduction</b>	<b>73</b>
<b>7.2</b>	<b>Methods</b>	<b>74</b>
7.2.1	Fuzzy Logic	74
7.2.2	Similarity Searches	75
7.2.3	GenDB	75
7.2.4	<i>Gramella forsetii</i> KT0803	75
7.2.5	The Reference Annotation	75
7.2.6	Running MicHanThi	76
<b>7.3</b>	<b>Algorithm</b>	<b>76</b>
7.3.1	Observation Preprocessing	76
7.3.2	Observation Evaluation and Selection	77
7.3.3	Annotation	78
7.3.4	Reliability of Annotations	80
7.3.5	Assignment of Additional Features	81
7.3.6	Comparing Annotations	82
<b>7.4</b>	<b>Design and Implementation</b>	<b>82</b>
<b>7.5</b>	<b>Results</b>	<b>83</b>
7.5.1	Evaluation of Human vs. MicHanThi Annotations	83
7.5.2	Evaluation of Revised Human vs. MicHanThi. Annotations	83
7.5.3	Evaluation of Revised Human vs. best BLAST and RAST Annotations	85
<b>7.6</b>	<b>Discussion</b>	<b>87</b>
7.6.1	Improving Annotation by Cross-Checking	87
7.6.2	Best BLAST Observation	87
7.6.3	RAST	87
7.6.4	Annotation of Hypothetical and Conserved Hypothetical Proteins	87
7.6.5	Annotation of Gene Functions	88
<b>7.7</b>	<b>Conclusions</b>	<b>89</b>

---

## Abstract

**Motivation:** The power of high-throughput sequencing technologies allows for the sequencing of microbial genomes and metagenomes on a routine basis. In contrast, accurate human annotation is still an elaborated and time consuming process. The MicHanThi annotation software was built to simulate the human annotation process to achieve a comparable accuracy in significantly less time.

**Results:** MicHanThi uses the results (observations) of different similarity search tools such as BLAST and InterProScan to predict the function for open reading frames (ORFs). To model the human reasoning process, MicHanThi uses Fuzzy Logic for the evaluation and selection of reliable observations.

The software was evaluated within the annotation jamboree of the marine bacterium *Gramella forsetii* KT0803. Compared to annotations created by the human annotators 71% of the annotations predicted by MicHanThi were syntactically identical. Additionally, 9% of the annotations were semantically equivalent.

The program performed best for *hypothetical* and *conserved hypothetical* genes, with or without *transmembrane region* and *signal peptide* predictions. Taking the results of MicHanThi into account, the manual annotation process was remarkably facilitated. In terms of speed and consistency MicHanThi clearly outperformed the human annotator as revealed by the subsequent manual cross-checking phase.

**Availability:** The software is freely available under the terms of the GNU General Public License through our web portal at <http://www.megx.net/michanthi>. An interactive browser for the visualisation of the results of this paper can be accessed at <https://gendb.mpi-bremen.de/gendb/CU207366>

**Contact:** fog@mpi-bremen.de

## 7.1 Introduction

Advancements in sequencing technology and the vast availability of sequencing capacity has lead to a dramatic increase of genomic and metagenomic sequence data. Today, a microbial genome can be sequenced in a matter of weeks. This time can be further reduced to days and hours when so called next generation sequencing technologies are applied (106).

Expert human annotation is claimed to be the most accurate approach for the functional annotation of genomes. Nevertheless, this process is extremely time consuming and lacks standardisation across annotation teams. To cope with the ever increasing amount of available genomic sequence information, several automatic annotation systems have been developed over the last years (107).

Two approaches for genome annotation can be distinguished: horizontal annotation, and vertical or subsystem annotation. In the horizontal annotation approach, the results of different tools and databases are used to predict a function for a single ORF. Neighbouring genes are normally not considered during the annotation process. Among others, systems using the horizontal annotation approach for the automatic prediction of gene functions are AutoFACT (33) and BASys (34).

Unlike horizontal annotation, vertical annotation predicts functions based on the conservation of ORFs in two or more genomes. Compared to the horizontal annotation, the order of genes is also important. Additionally, *subsystems* can be used to enrich the annotation process. Subsystems are groups of genes commonly based on metabolic pathways but they can be any expert-defined group of functionally related genes. A gene family is created for each gene in a subsystem. New sequences are then compared to profiles representing these families. Ergo (35) was the first annotation system to facilitate the vertical annotation approach. Today, commonly used systems for the automatic annotation of genomes are the IMG system (6) and the RAST system (5).

All systems incorporate tools for ORF calling, similarity search tools, and tools for the prediction of gene functions. According to the authors, few of these systems can be installed locally and only limited tool configuration parameters are available using web based systems. (107) propose the development of custom annotation pipelines if the complete control of the annotation process is needed or if specific problems need to be addressed.

To build custom annotation pipelines all tools need to be flexible regarding the source and type of input data as well as the output format. Ideally, they should focus on single tasks to easily compare and select the tool / tool combination which best suits the requirements at hand.

In the human annotation process, the integration of heterogeneous information, e.g., provided by different tools is highly complex. Even if fixed annotation rules exist, the weighing of the different tools and databases varies for the final decision on the predicted function of a gene. Thorough knowledge of the data involved is a key asset for the human annotator. It must be clear that a full representation of such domain knowledge can never be achieved in an automatic annotation system. However, a formal description of the annotation process is still necessary.

Fuzzy Logic can be used to represent the evaluation of observations. Fuzzy Logic and fuzzy sets offer flexible reasoning abilities by regarding precise reasoning / precise logic as a boundry case (108). Rather than assigning an element to a set or not (true / false decisions), it assigns a proportional truth value representing an element's degree of membership to a set. As such, an element may belong to multiple and even contradicting sets at the same time.

MicHanThi was build for the horizontal annotation of genes applying fuzzy logic. It uses abstraction layers to completely encapsulate the loading and the writing of data. It is designed to be easily expandable integrating new similarity search tools, and it offers extensive control over the parameters involved in the decision making process. The annotation process is divided into three main steps: the evaluation and selection of observations, the creation of annotations, and the evaluations of the created annotations.

The software was successfully used in several genome annotation projects. It was comparatively evaluated against the expert human annotation of *Gramella forsetii* KT0803 (61), as well as the automatic approaches of transferring the *best BLAST observations* and annotations created by the RAST system.

## 7.2 Methods

### 7.2.1 Fuzzy Logic

*Linguistic variables* are a key concept of fuzzy logic. Instead of values they consist of words or sentences taken from a natural or synthetic language. Rather than relying on crisp / fixed thresholds each value represents a function (*membership function*) defining the degree of membership of an element to the different sets represented by the linguistic variable. The definition of the membership functions is based on the knowledge of a domain expert.

Let the terms *evaluate* and *coverage* be linguistic variables describing a BLAST hit. Further let 'unreliable', 'uncertain', 'reliable', and 'very reliable' be possible values of *evaluate* and 'none', 'partial', and 'complete' be values of *coverage*. Let the linguistic variables *good*, *bad* be attributes describing the reliability of an observations. A rule describing a BLAST observation could be:

```
IF evaluate is reliable AND coverage is complete
THEN hit is good
```

The rule is evaluated by *fuzzifying* the numerical values of *evaluate* and *coverage*. A value is fuzzified by applying the membership function of a linguistic variable to a value



and thus determining its degree of membership. The operator AND is used to combine the values of evaluate and coverage. Last, the result of the rule is *defuzzified*. This step converts the linguistic variable *good* back into a numerical value which can be used to rank the observations.

### 7.2.2 Similarity Searches

MicHanThi uses the result (*observations*) of standard prediction tools to generate an annotation for the sequence of interest (query sequence). The following tools and databases were used in this work:

- BLAST (31) vs. NCBI nr; Blast searches were performed against the NCBI nr database (including Swiss-Prot(47))
- InterProScan (8) (vs. InterPro databases Pfam / TIGRfams); From the results obtained by running InterProScan against the InterPro database (11), only those results that are based on Pfam (52) and TIGRfams (53) entries are considered.
- SignalP (62), TMHMM (63); SignalP-HMM and TMHMM are two hidden Markov model-based tools that predict protein targeting and transmembrane helices of a protein. In the annotation of *Gramella forsetii* KT0803 the results of these tools were used to verify functional annotations and to assign additional features to *hypothetical* and *conserved hypothetical* proteins.

### 7.2.3 GenDB

The GenDB system (9) was used as a framework for running the different tools. It is build around a relational database management system and uses distributed cluster computing to generate a set of observations for each query sequence. Furthermore, it offers a web interface allowing the user to annotate ORFs manually and inspect the results of the automatic annotation process. The software system is freely available at: <http://sourceforge.net/projects/gendb/>

A comprehensive discussion about different annotation systems and why the GenDB annotation system was chosen as a framework can be found in (10).

### 7.2.4 *Gramella forsetii* KT0803

*Gramella forsetii* KT0803 (61) is a marine flavobacterium which has been isolated from North Sea surface waters during a summer phytoplankton bloom. It was the first marine representative of the phylum *Bacteroidetes* to have its genome sequence completely determined.

### 7.2.5 The Reference Annotation

After gene prediction using the tool MORfind (Waldmann, unpublished) the non-redundant list of ORFs (3593) was imported into the GenDB annotation system and the different tools to functionally characterise the ORFs were run. Approximately 1.4 million observations were created.

Based on these observations the manual annotation was created by a collaborative work of an expert annotation team. In total 15 people (eleven trainees and four senior annotators) created a first draft annotation during nine weeks.

Following the initial annotation, each annotation was evaluated by an expert to achieve the highest accuracy possible, adding missing information as well as creating new annotations if necessary. This evaluation phase required three more weeks. Additionally, several months were spent to investigate the physiological and metabolic capabilities of the organism.

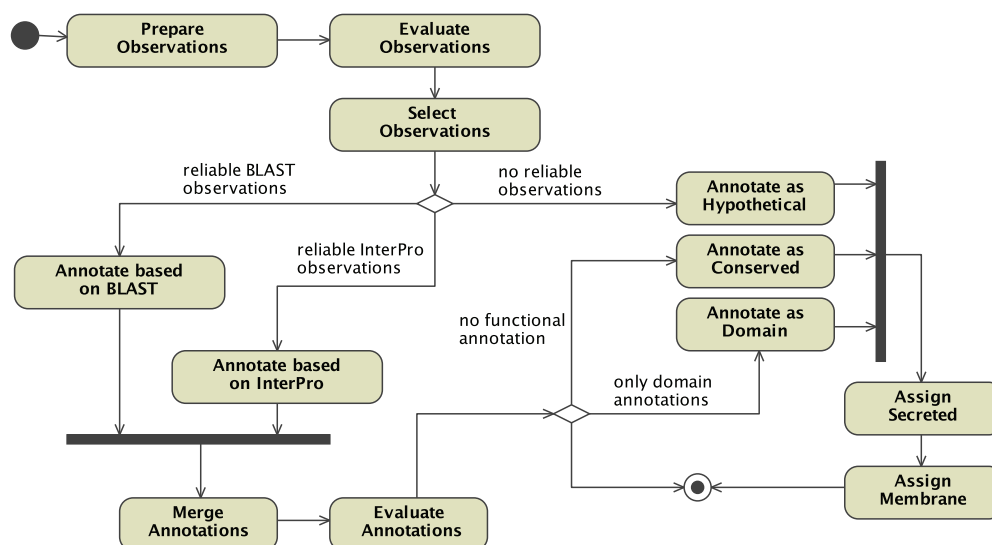


Figure 7.1: The Annotation Process.

## 7.2.6 Running MicHanThi

72 instances of MicHanThi were run in parallel on 18 compute nodes. Each compute node featured two hyper-threading Xeon CPUs clocked at 2.80 GHz each. Additionally, each compute node was equipped with a total 4 GB of main memory. On this set-up the annotation of the 3593 ORFs predicted in the *Gramella forsetii* KT0803 genome took roughly eight minutes. Overall, the prediction of ORFs, the similarity searches, and the prediction of gene functions took less than 12 hours. Of these 12 hours, MicHanThi ran for approximately 20 minutes.

## 7.3 Algorithm

Figure 7.1 shows an overview of the steps involved by MicHanThi to predict a gene function.

### 7.3.1 Observation Preprocessing

Descriptions of BLAST observations may contain many different types of information like: the gene function, a gene name, an EC number, an organism name, and synonyms, among others. An initial step when creating an annotation for a gene is to separate these pieces of information.

Additionally, an entry in NCBI nr may contain more than one functional description, separated by the entry's source database identifier. Splitting the description line into single descriptions is necessary for the annotation process. An extended Backus-Naur form as used by the W3C to specify XML<sup>1</sup> is used to formally describe the syntax of valid description lines.

```

ENTRY ::= (DEB | PDB | PIR | PRF | REF | SP)+
DEB ::= (GI)? ("dbj" | "gb" | "emb") "|" L (L | D)+ "." D
      "|" " FUNC " [" ORG "]
PDB ::= (GI)? "pdb|" (L | D)+ "|" " FUNC (FROM)?
PIR ::= (GI)? "pir|" L (L | D)+ FUNC
      ("("ECNUM")")? (DBREF)? ("-" ORG)?
PRF ::= (GI)? "prf|" L (L | D)+ FUNC [" ORG "]
  
```

<sup>1</sup>The Specification of the *Extensible Markup Language* (XML) can be found at <http://www.w3.org/TR/REC-xml/>

```

REF      ::= (GI)? "ref|" L L "_" D+ "." D "|" " FUNC "[" ORG "]"
GI       ::= "gi|" D+ "|"

FUNC     ::= "functional description of the protein" (GENENAME)? (ECNUM)?
FROM     ::= "From " ORG ", additional information"
DBREF    ::= "database identifier"
ORG      ::= "name of the organism containing this protein"
GENENAME ::= "the gene name of a protein"
ECNUM    ::= ("EC ")? D "." D+ "." D+ "." D+ |
            ("EC ")? D "." D+ "." D+ "." "-" |
            ("EC ")? D "." D+ "." "-" |
            ("EC ")? D "." "-" |
D        ::= [0-9]
L        ::= [a-zA-Z]

```

Descriptions from Swiss-Prot entries contain more information than other databases. Especially, the list of possible synonyms of a given function is important for the subsequent annotation process.

```

SP       ::= SPID (CLEAVED | MULTI | SINGLEFUNC | RARE)
SPID     ::= (GI)? "sp|" L(L D)+ "|" L+

CLEAVED  ::= PRECURSOR " [Contains: " FUNCLIST "]"
PRECURSOR ::= "name of the precursor protein" (SYNONYM)*
FUNCLIST ::= SINGLEFUNC (" ; " SINGLEFUNC)+
SINGLEFUNC ::= FUNC (SYNONYM)*
SYNONYM  ::= "(" FUNC | ECNUMBER ")"

MULTI    ::= MULTIFUNC | BIFUNC
MULTIFUNC ::= SINGLEFUNC " [Includes: " FUNCLIST "]"

BIFUNC   ::= SINGLE " [Includes: " FUNCLIST "]"
SINGLE    ::= BFUNC (" (" SYNONYM ")")*
BFUNC    ::= "Bifunctional protein" GENENAME
RARE     ::= SINGLEFUNC " [Includes: "FUNCLIST"] [Contains: "FUNCLIST]"

```

Another important aspect of the observation preprocessing is the discarding of uninformative terms. A term is considered to be uninformative if it does not describe a gene function or if it is not part of a functional description (33). Examples are locus tags internally used by genome annotation projects, words like ‘and’, ‘or’, ‘found’, and ‘in’, among others. Terms like *putative*, *possibly involved in*, and *like* are considered to be qualifiers weakening an annotation. If a BLAST observation is reliable but does not contain any informative terms it is considered to be a protein of unknown function.

No preprocessing is needed for InterPro, SignalP, and TMHMM observations because the description and the vocabulary used are well defined.

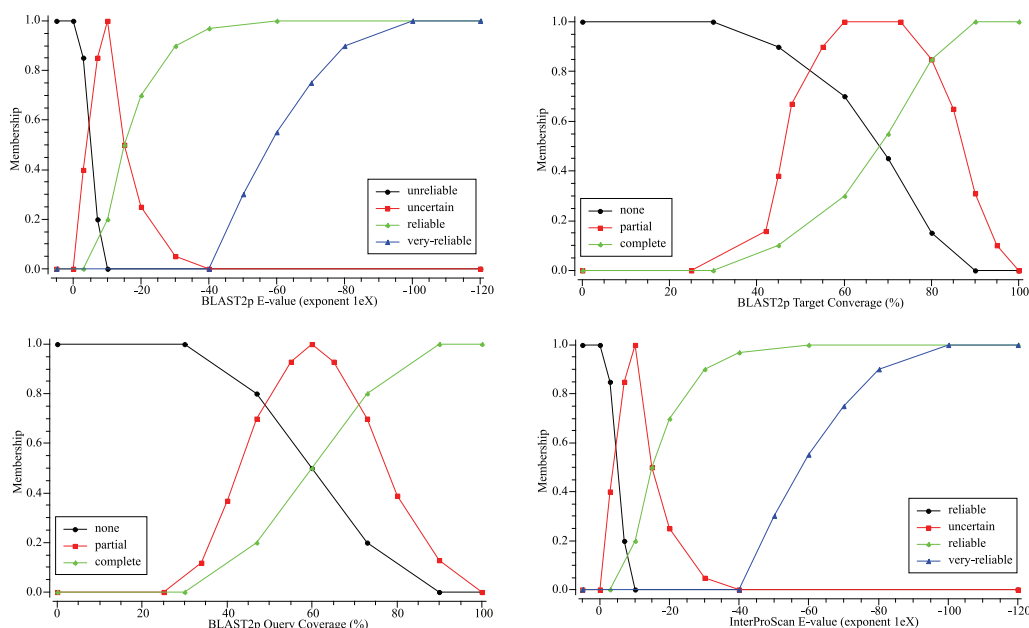
### 7.3.2 Observation Evaluation and Selection

Next in the annotation process is the evaluation and selection of reliable observations for creating the annotation. MicHanThi uses fuzzy logic to express the reliability of an observation. Based on the tool that created the observations different methods for the evaluation have to be considered.

#### BLAST Observations:

A long list of potentially related proteins is found if BLAST matches a query sequence against the NCBI nr database. Each observation within this list has to be evaluated to find those that are most likely orthologous to the query sequence. To support the reasoning process BLAST offers four criteria: the coverage of the alignment (query / target), the number of identical bases and positive substitutions, the bit and raw scores, and the E-value.

MicHanThi uses the alignment coverage and the E-value to evaluate BLAST observations. The coverage is evaluated separately for the observation (target) and for the query sequences. Three linguistic variables are used to express the reliability of an BLAST observation: *evaluate*, *coverageORF*, and *coverageDB*. Each variable may have up to four values: *coverageDB/ORF* (none, partial, complete), and *evaluate* (unreliable, uncertain, reliable, very reliable). The membership functions for each criteria are depicted in Figure 7.2.



**Figure 7.2:** Definition of the member functions for the linguistic variables of BLAST and InterProScan.

Based on a rule base comprising 36 rules, an observation is assigned to one of the four categories: *bad*, *average*, *good*, and *very-good*. Of all observations created by BLAST, the 25 best but all very-good observations are selected for the annotation process.

### InterPro Observations:

For *hidden Markov model* based InterPro results only the E-value is important to select an observation (Figure 7.2). Four simple rules are defined by the rule base. Each possible value of the linguistic variable (*very-reliable*, *reliable*, *uncertain*, and *unreliable*) is assigned to one of the four reliability classes (*very-good*, *good*, *average*, and *bad*). All observations describing a Pfam or TIGRFams family or domain, except bad observations, are kept for the annotation process.

### SignalP / TMHMM Observations:

Results returned by SignalP and TMHMM can be interpreted as simple *true* / *false* statements representing the presence or absence of certain features.

TMHMM observations predicting transmembrane helix regions are always kept, other TMHMM observations are discarded.

### 7.3.3 Annotation

Annotations for observations based on InterPro and BLAST are created separately. Later these annotations are checked and merged if they describe the same function. Annotations whose reliability is less than 85% of the most reliable annotation are deleted.

If an annotation is supported by either InterPro or Swiss-Prot observations then the annotation is deleted only if their reliability is less than 70% of the most reliable annotation. Additionally, if two annotations are equally good then annotations not based on Swiss-Prot or InterPro are deleted. Annotations based on InterPro and Swiss-Prot

are favoured over other annotations because these databases are manually curated and generally assumed to be of high quality.

If no observations were reported, then an ORF is annotated as *hypothetical protein*. Such ORFs can further be distinguished into the following classes.

- *hypothetical protein*: An ORF is described as a *hypothetical protein* if no matches could be found in any of the sequence databases or if matches exist but they are considered unreliable ( $E - value \geq 1e^{-3}$ ).
- *conserved hypothetical protein*: The attribute *conserved* is assigned to a hypothetical ORF if at least one reliable match could be found in one of the sequence databases. A match is considered reliable if  $E - value < 1e^{-3}$  and ORF coverage  $\geq 30\%$  and DB coverage  $\geq 30\%$ .
- *protein containing*: If no reliable BLAST observations were found but a reliable observation describing an InterPro domain has been found, then an ORF is annotated as a *protein containing domain*.
- *transmembrane prediction*: For ORFs that have at least two reliable transmembrane helix predictions the attribute *membrane* is assigned.
- *signal peptide prediction*: If no more than one transmembrane helix was predicted for an ORF and a reliable signal peptide prediction exists, then the ORF is annotated as *secreted*. A signal peptide prediction is considered to be reliable if its probability as reported by SignalP-HMM is  $\geq 0.75$  and its cleavage site probability is  $> 0.5$ .
- *transmembrane and signal peptide predictions*: If exactly one transmembrane helix prediction exists for an ORF and the predicted signal peptide prediction is uncertain because its HMM cleavage site probability is  $\leq 0.5$ , then the ORF is annotated as *membrane or secreted*.

### InterPro:

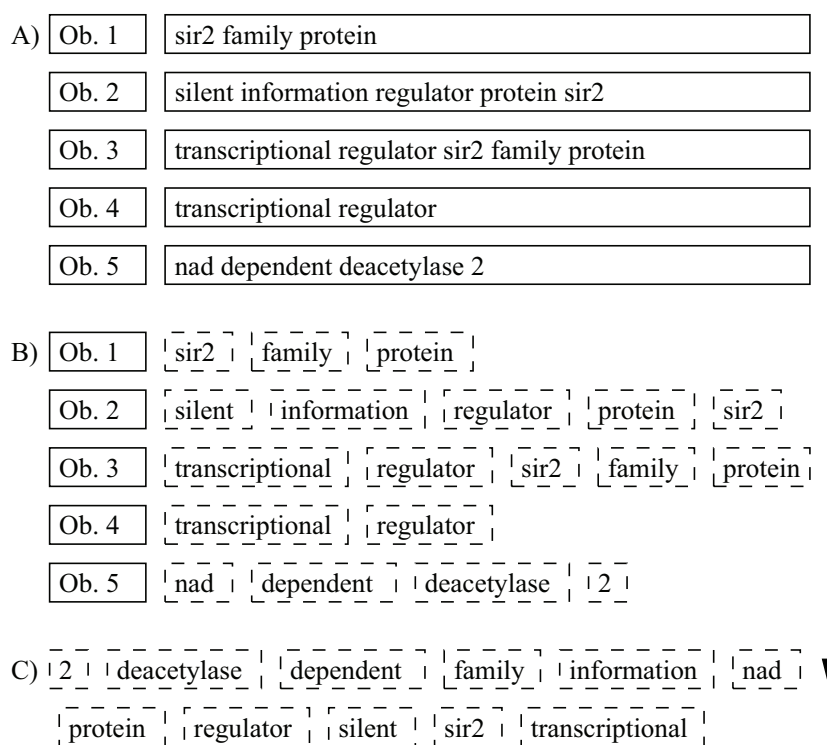
InterPro entries include information about relations to other InterPro entries. An entry might be a *parent* or a *child* of another entry. An entry may *contain* another entry or in case of a domain entry it might be *found in* another entry. Before annotations based on InterPro are created these relations are checked. Parent entries are deleted if the child is at least 80% as reliable as the parent. InterPro entries found in other entries are also deleted if the container is at least 80% as reliable as the contained entry. For each remaining InterPro observation, an annotation is created.

### BLAST:

Often, the list of BLAST observations includes different functions as well as a variety of descriptions of the same function. The five observations shown in Figure 7.3 A describe the same function with different specificity. Observation four uses the most general description. Observations one through three differ only in wording and observation five uses a synonym to describe the function. To assign a function to an ORF, MicHanThi tries to find the common denominator.

To find the common denominator among all BLAST observations, the observations have to be grouped. A group is a tuple of a number of words present in a description (*atoms*) and all observations containing that particular combination of words (*support*). The order of a group is the number of included *atoms* and its reliability is the average reliability of all supporting observations.

A group may be a subset of another group if all its atoms as well as all its supporting observations are contained in a group of an higher order. Groups are considered to be



**Figure 7.3:** Splitting observation descriptions into atoms. Sample observations are taken from *orf1081* of the *Gramella forsetii* KT0803 annotation project.

invalid if they either are not supported by any observation or if they contain an atom multiple times. Invalid groups are deleted.

All descriptions are split and a non-redundant list of *atoms* is created (Figure 7.3 B and C). Initially, for each atom a group of order 1 is created (Figure 7.4 first row). Groups of order  $n+1$  are created through the combination of groups of order  $n$  (with  $n \geq 1$ ) and groups of order 1 (Figure 7.4 second and third rows). Observations containing the new combination of atoms are copied to the new group. This process is iteratively repeated until no valid groups of order  $n+1$  can be constructed.

For each remaining group an annotation is created based on its most reliable supporting observation. If a group contains Swiss-Prot observations then the functional description is based on this observation.

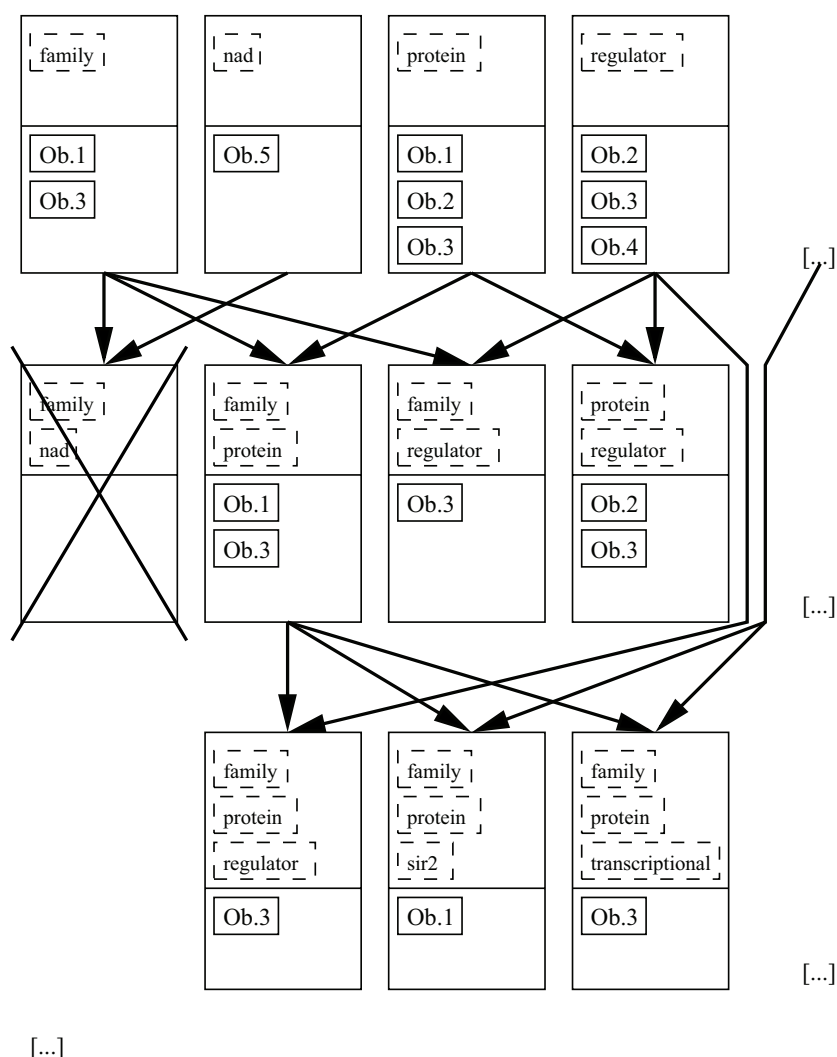
### 7.3.4 Reliability of Annotations

Three quality attributes are assigned to each annotation. The first two attributes are the reliability of the best supporting observation and the average observation reliability. The third attribute is a measurement of how closely an annotation resembles an observation. Since annotations based on BLAST observations may not contain all words used in an observation's description this attribute is necessary to describe the annotation's accuracy.

The accuracy of annotations not based on BLAST observations is always 1.0. For annotations based on BLAST observations it is the ratio of informative terms used in the observation's description and informative terms used by the annotation. If all informative terms are used then the annotation's accuracy is 1.0.

The accuracy of an annotation (*glycosyl hydrolase*) based on the observation *glycosyl hydrolase, family 32* and *glycosyl hydrolase, family 53* would be 0.5.

Each annotation is also assigned an overall reliability. This reliability expresses the annotation's accuracy and the best observation's reliability. If an annotation is considered



**Figure 7.4:** *Generation of groups. Each group contains a list of atoms and all observations containing each atom in its description.*

*uncertain* or *unreliable* then its gene function is prefixed with *[similarity to]* or *[weak similarity to]* respectively.

### 7.3.5 Assignment of Additional Features

Further features like EC numbers, gene names, and a list of GO numbers are annotated if supporting observations are found. If an annotation is based on BLAST observations, then the supporting observations are checked for matches against the Swiss-Prot database. If a Swiss-Prot entry is found, the EC number and gene name are extracted, as well as associated GO numbers.

EC and GO numbers are also assigned if the annotation is based on an InterPro observation. Gene names may be found in the InterPro entry's description field. This is a free text field and extracting the gene name from this field is almost impossible. Therefore, MicHanThi does not assign gene names based on InterPro observations.

### 7.3.6 Comparing Annotations

The annotation created by the human expert team was considered to be the reference annotation. The comparison between MicHanThi and the human annotation was done automatically.

To determine if two annotations describe the same function the product field was syntactically compared. Two annotations are considered to be equal if all terms used in annotation A are also used in annotation B. The order of terms was not considered. Two annotations are different if at least one character differs between their products. This includes cases in which one product describes the function in plural and another in singular form. Also, product descriptions containing spelling mistakes are considered to be different.

Another problem that arises when annotations are automatically compared is that the specificity of the description may differ. An ORF may simply be annotated as *glycosyl hydrolase* while a more precise description could be *glycosyl hydrolase, family 53. Subset matches* were introduced to accommodate this fact. Annotation D is considered to be a subset match if it uses a subset of terms used in annotation C and vice versa.

Thorough investigation of ORFs of unknown function was chosen because strict annotation rules exist.

The *precision* of an automatic approach describes the approach's ability to create only correct annotations. Since annotations created by a human expert are considered to be correct in all cases, the automatic approach has to create these annotation as well (*true positives*). Also no additional annotations should be created (*false positives*). Therefore, the term precision can be defined as the number of matching annotations, divided by the number of all automatically created annotations in a certain sub class.

$$precision(\%) = \frac{true\ positives}{true\ positives + false\ positives} * 100 \quad (7.1)$$

$$\leftrightarrow precision(\%) = \frac{\#\ matches\ human\ to\ automatic}{\#\ automatic\ annotations} * 100 \quad (7.2)$$

The term *recall* is the capability to reproduce all annotations which have been created by the human annotator. It can be expressed as the number of matches divided by the number of matches plus the number of those annotations which were not created by the automatic approach (*false negatives*). Again only annotations in a certain category are considered.

$$recall(\%) = \frac{true\ positives}{true\ positives + false\ negatives} * 100 \quad (7.3)$$

$$\leftrightarrow recall(\%) = \frac{\#\ matches\ human\ to\ automatic}{\#\ total\ human\ annotations} * 100 \quad (7.4)$$

## 7.4 Design and Implementation

MicHanThi uses an abstract description of the sources of information, such as the annotation system, as well as the analysis tools. It consist of four modules: (i) the IO module (ii) the DATA module, (iii) the TOOLS module, and (iv) the ANNOTATOR module.

### Programming Language and External Libraries

The software is implemented in version 1.5 (v5) of the Java programming language. It depends on four external libraries which need to be part of the Java CLASSPATH environment: JSAP, mbfuzzit, MySQL Connector/J, xerces-java. The libraries JSAP (<http://www.martiansoftware.com/jsap/>) and xerces-java (<http://xml.apache.org/>) are used to parse configuration files written in XML and merge those options with options obtained from the command line. To connect to the MySQL server software, the MySQL Connector/J (<http://www.mysql.com>) which is used by GenDB to store the data. For the fuzzy logic reasoning engine the mbfuzzit (<http://mbfuzzit.sourceforge.net/>) library is used.



**Table 7.1:** Overall statistics of the comparison of annotation created by human annotators and annotations created by MicHanThi.

number of ORFs: 3593		
	#annotations	per ORF
human annotations:	12080	3.4
automatic annotations:	5007	1.4
	#annotations	% of ORFs
exact matches:	1486	41.4%
subset matches of automatic annotation:	311	8.7%
subset matches of human annotation:	315	8.8%
overall	2112	58.9%

## 7.5 Results

The manually annotated genome of *Gramella forsetii* KT0803 provided a basis to evaluate the performance of MicHanThi. Each annotation created for an ORF by the software was compared to the annotation created by the human expert.

Additionally, the performance of MicHanThi was compared to the approach *annotation by transferring the best BLAST observation* and annotations created by the RAST system.

### 7.5.1 Evaluation of Human vs. MicHanThi Annotations

In the first evaluation phase the results produced by MicHanThi were compared to the unrevised human created annotations. Overall, 2115 annotations of the 3593 ORFs (59%) matched (table 7.1). Of these, 17% are subset matches where either the annotation created by the human annotator or by MicHanThi is more specific than the other annotation.

Table 7.2 (column ‘H↔M’) shows detailed statistics about the classes of annotations without a functional assignment. The number of annotations created by the human annotator and by MicHanThi is about the same in all classes. Nevertheless, the number of matching annotations is low (overall 49% precision and 53% recall). The largest differences are within the assignment of the attributes *membrane* and *secreted*. In most cases MicHanThi assigns these attributes to more ORFs than the human annotator does.

### 7.5.2 Evaluation of Revised Human vs. MicHanThi. Annotations

The second evaluation step was the comparison of the annotations created by MicHanThi and the annotations created by the human annotator once these annotations were revised by experts. Annotations created by MicHanThi were the same in both cases. The overall number of matches increased by approximately 12% from 59% to 71% (table 7.3), including 18% of subset matches.

Taking a closer look at the statistics for the classes without functional assignment (table 7.2 column ‘H↔M’) shows that the precision and recall values increased in all cases. The overall value for the precision rose to 72% and the overall recall value is 73% compared to 50% and 52% as seen before. The increased performance can particularly be seen in class *hypothetical*. Within this class the recall value ranged from 92% to 98%. Notably, the correct assignment (precision) of the attributes *transmembrane* and *secreted*

**Table 7.2:** Detailed comparison of annotations for ORFs without a functional assignment. Column H represents the number of annotations created by the human annotator in each category before the annotations were revised by experts. Column Hr holds the number after the annotations were revised. The number of annotations created by MicHanThi can be found in column M and those created by RAST are in column R. The comparison of unrevised human annotations and the automatically created annotations can be found in columns ‘H→M’. The column groups ‘Hr→M’ and ‘Hr→R’ compare the revised, manually created annotations to the automatic approaches MicHanThi and RAST.

ORFs without function	H	Hr	M	R	matches H→M		matches Hr→M		matches Hr→R	
					Total %ORFs	[Precision, Recall]	Total %ORFs	[Precision, Recall]	Total %ORFs	[Precision, Recall]
	1598	1615	1636	1467	826	23%	1184	33%	728	20%
						[50%, 52%]		[72%, 73%]		[50%, 45%]
<b>Hypothetical</b>										
protein	596	465	510	1147	425	12%	455	13%	458	13%
membrane protein	95	99	114	65	50	1%	94	3%	64	2%
secreted protein	107	171	205	81	63	2%	158	4%	77	2%
membrane or secreted protein	0	10	10	19	0	0%	6	0%	3	0%
						[83%, 71%]		[89%, 98%]		[40%, 99%]
						[44%, 53%]		[82%, 95%]		[98%, 65%]
						[31%, 59%]		[77%, 92%]		[95%, 45%]
						[0%, -]		[60%, 60%]		[16%, 30%]
<b>Conserved</b>										
conserved protein	443	341	266	3	185	5%	205	6%	2	0%
conserved membrane protein	64	107	78	36	26	1%	63	2%	36	1%
conserved secreted protein	84	167	135	42	33	1%	94	3%	36	1%
conserved membrane secreted protein	8	12	7	0	0	0%	5	0%	0	0%
						[70%, 42%]		[77%, 60%]		[67%, 1%]
						[33%, 41%]		[81%, 59%]		[100%, 34%]
						[24%, 39%]		[70%, 56%]		[86%, 22%]
						[0%, 0%]		[71%, 42%]		[-, 0%]
<b>Domain</b>										
protein containing membrane protein containing secreted protein containing membrane or secreted protein containing	178	160	195	38	36	1%	59	2%	25	1%
	8	47	67	19	3	0%	34	1%	18	1%
	14	35	47	15	5	0%	13	0%	9	0%
	1	1	2	2	0	0%	0	0%	0	0%
						[18%, 20%]		[30%, 37%]		[66%, 16%]
						[4%, 38%]		[51%, 72%]		[95%, 38%]
						[11%, 36%]		[28%, 37%]		[60%, 26%]
						[0%, 0%]		[0%, 0%]		[0%, 0%]

**Table 7.3:** Overall statistics of the comparison of the revised human created annotations and annotations created by MicHanThi.

number of ORFS: 3593		
	#annotations	per ORF
human annotations:	13938	3.9
automatic annotations:	5007	1.4
	#annotations	% of ORFs
exact matches:	1894	52.7%
subset matches of automatic annotation:	338	9.4%
subset matches of human annotation:	306	8.5%
overall	2538	70.6%

increased by 38% and 46% percent points, respectively. The increased performance in the other two classes is still noteworthy and will further be discussed in section 7.6.4.

In class *hypothetical* 32 annotations created by the human experts did not match those annotations created by MicHanThi for the same ORF. Sixteen of these mismatches had differences in the assignment of the attributes *transmembrane* and *secreted*. MicHanThi considered 11 ORFs to be *conserved*, assigned a function to four ORFs, and a domain to one (table 7.4 A).

A total of 127 annotations created by MicHanThi in the same category did not match those annotations created by the human annotator. Of these mismatches, 78 ORFs were classified as *conserved* by the human annotator. Additionally, 11 annotations include the information that the ORF contains a functional domain as described by Pfam or InterPro. A function was assigned to 22 ORFs by the human annotator and 16 annotations differed in the attributes *transmembrane* and *signal peptide*.

Table 7.4 shows the different types of mismatches for the classes *conserved* and *domain*. The first part (A) of this table describes mismatches to the human annotations and the second part (B) describes mismatches to annotations created by MicHanThi.

### 7.5.3 Evaluation of Revised Human vs. best BLAST and RAST Annotations

To show the performance of MicHanThi with respect to other automatic approaches, the revised human annotations were compared to those created by transferring the *best BLAST observation* and the annotations created by the RAST annotation system.

For the approach *best BLAST*, only 30% of the annotations matched. This approach left 1305 ORFs without a functional assignment and had low precision (35%) and recall (29%) values in this class of annotations. Also, 65% of the annotations automatically created in these classes were incorrect, an exception can be found in class *hypothetical*. In this class almost all (98%) annotations created by the human annotator were also created by the computer. A precision of only 35% in this class indicates a large number of additionally created annotations.

The results obtained from the RAST system matched the human annotations in approximately 50% of the cases. To 1467 ORFs no function was assigned. The precision of the RAST server in this class was 50% and the recall value was 45%. RAST uses the vertical annotation approach and assigned 845 (24%) ORFs to subsystems.

**Table 7.4:** *Details of the mismatches in classes hypothetical, conserved, and domain. A) mismatches to the human annotations. B) human annotations that do not match annotations created by MicHanThi.*

A) human vs. MicHanThi annotations		B) MicHanThi vs. human annotations	
Type of Mismatch	Count	Type of Mismatch	Count
1) class hypothetical attribute	16	1) class hypothetical attribute	16
2) class hypothetical vs. class conserved	11	2) class hypothetical vs. class conserved	78
3) class hypothetical vs. class domain	1	3) class hypothetical vs. class domain	11
4) class hypothetical vs. class function	4	4) class hypothetical vs. class function	22
5) class conserved attribute	15	5) class conserved attribute	15
6) class conserved vs. class hypothetical	76	6) class conserved vs. class hypothetical	11
7) class conserved vs. class domain	12	7) class conserved vs. class domain	27
8) class conserved vs. class function	157	8) class conserved vs. class function	66
9) class domain attribute	72	9) class domain attribute	72
10) class domain vs. class hypothetical	11	10) class domain vs. class hypothetical	1
11) class domain vs. class conserved	27	11) class domain vs. class conserved	12
12) class domain vs. class function	106	12) class domain vs. class function	173

## 7.6 Discussion

### 7.6.1 Improving Annotation by Cross-Checking

After the initial evaluation phase, annotations created by MicHanThi matched those created by the human annotator in 58% of the cases (table 7.1). The number of matches increased to 71% once each manually created annotation was checked by an expert, correcting wrongly created or imprecise annotations. Particularly, in the class of annotations where a functional assignment was not possible, the number of matching annotations was initially low (table 7.2 column ‘H ↔ M’). This was unexpected because especially within this class the computer should be able to achieve a large number of correctly created annotations because strict and simple rules exist.

The reason that this does not emanate from the comparison can foremost be found in the inconsistencies among the annotations created by the human annotators. This becomes apparent when the annotations before and after the revision phase are compared to each other (table 7.2 columns H and Hr). More than 1800 annotations were corrected by the human experts. In each class the number of annotations differ by up to 60%. For example, before the annotations were revised 107 ORFs in the class *hypothetical* were assigned the attribute *secreted*. After experts corrected the annotations, *secreted* was assigned to 171 ORFs. The number of matching annotation increased from 63 to 158. Accordingly, MicHanThi’s precision increased from 31% to 77% and its recall from 59% to 92%.

### 7.6.2 Best BLAST Observation

Differences in the human created annotations and those created by transferring the best BLAST observation could be explained by not looking at the list of all BLAST observations for that particular ORF. The best match often does not reflect all available information about the protein because this information could have changed over time and the publicly available annotation may not have been updated accordingly. Also, if no function could be assigned to the ORF then the created annotation was imprecise due to the missing assignment of additional attributes like *transmembrane* and *secreted*.

### 7.6.3 RAST

Vertical annotation offers the advantage that a function may be proposed for otherwise *hypothetical proteins*. If a subsystem is conserved in an organism only missing few genes and these genes are covered by *hypothetical proteins* a functional relation may be assumed. In case of the genome of *Gramella forsetii* KT0803, RAST associated 25 *hypothetical proteins* with subsystems offering the human annotator additional information about the gene neighbourhood.

The results created by MicHanThi are based on the same observations the human used to annotate the genome. The RAST annotations referred to in this comparison were created in January 2009. Therefore, newer versions of the databases were used, possibly including *Gramella forsetii* KT0803 annotations. While the bias on the assignment to a subsystem can be neglected, it may pose a problem for results obtained by BLAST.

### 7.6.4 Annotation of Hypothetical and Conserved Hypothetical Proteins

MicHanThi performed well in class *hypothetical* where its performance in the classes *conserved*, and *domain* is not as good. In most cases in which the annotations of an ORF do not match either the human annotator (table 7.4 B.8 – 67 ORFs), or the computer (table 7.4 A.8 – 157 ORFs) assigns a function to the ORF.

As an example, ORF orf279 was annotated as *1,4-alpha-glucan branching enzyme* by the human annotator. This annotation was based on 11 BLAST observations with E-values between  $1e^{-3}$  and  $2e^{-16}$  of which four observations had a target coverage of less than 25%. Eight of the 11 BLAST observations did not describe any function. One observation described a *isoamylase N-terminal domain protein* ( $2e^{-15}$ , good alignment coverages) but the corresponding Pfam observation had to be considered unreliable ( $5e^{-2}$ ). Two observations described a *1,4-alpha-glucan branching enzyme* (one being automatically derived by querying the COG database). There were no reliable InterPro observations and only two very weak Swiss-Prot observations ( $1e^{-4}$ , target alignment coverage less than 20%). It is left to the reader to decide whether this ORF should be annotated as *1,4-alpha-glucan branching enzyme* or if it should be annotated as *conserved hypothetical protein* instead.

For most of the mismatches for which the human annotator predicted a function and MicHanThi did not, the list of observations was alike. In some cases reliable observations have been found but these were hits to proteins of unknown function.

During the manual annotation process the genome of *Gramella forsetii* was divided into sections of approximately 250 ORFs. Interestingly, the number of mismatches (computer *conserved* vs. human *function*) was highest in those sections that were initially annotated by trainees. In these sections, up to ten mismatches were found for the annotations of 250 ORFs. In sections annotated by more experienced biologists, the number of mismatches was less than four. MicHanThi predicted functions for 157 ORFs considered to be *conserved hypothetical*. Most of these annotations had low reliability and accuracy values and could, therefore, be spotted easily.

Seventy six ORFs annotated by the human annotator as *conserved hypothetical protein* (table 7.4 A.6) were annotated as *hypothetical protein* by MicHanThi. In all cases, only very few observations having an E-value of less than  $1e^{-3}$  could be found. Almost all of these annotations had a low query coverage or a low target coverage (the alignment created by BLAST covers less than 20% of the query / target). Eleven automatically created *conserved hypothetical protein* annotations (table 7.4 B.6) were not matched by the human annotator. Four of these annotations are based on weak observations as the annotations described above. The remaining seven annotations are based on reliable observations with E-values down to  $1e^{-179}$  and alignments covering the query as well as the target sequences in most cases.

Across all classes, mismatches could be explained by the assignment of the attributes *membrane* and *secreted*. This can particularly be seen in class *domain* (Table 7.4 A.9 and B.9). According to the very strict rules described in Section 7.3.6 MicHanThi was correct in all cases.

### 7.6.5 Annotation of Gene Functions

The diversity of terms used in BLAST observations to describe the same biological function poses a significant problem for the automatic evaluation of annotations. In case a computer needs to interpret these terms, they would only be equal if they are spelled exactly the same way. Subset relations between two or more observations might be established if annotations are described by more than one term. These subsets may then be used to infer a common function with different specificity.

An illustration of this problem can be found in table 7.5. This list of annotations is an excerpt of annotations from the first one hundred ORFs in the *Gramella forsetii* genome. It shows the annotations created by the human annotator (Hr) compared to those created by MicHanThi (A). The annotations for each ORF describe the same gene function using different words. The difference in wording ranges from subtle (orf10) to *obvious* (orf62).

Most differences in this list can be explained by the observations that were used for the annotation. In most cases the human annotator used Pfam, while the computer used InterPro (orf10, orf40) or Swiss-Prot entries (orf33, orf58, orf95). Even though the

**Table 7.5:** Annotations created by a human annotator (Hr) and automatically created annotations (A) for the same ORF that use semantically equivalent / similar descriptions for the gene product. These annotations are wrongly assumed to be mismatches by the semi-automatic evaluation process.

ORF	(H)uman / (A)utomatic Annotations
orf7	Hr sensor histidine kinase / response regulator hybrid, sugar binding
	A two-component system sensor histidine kinase / response regulator, hybrid ('one component system')
orf10	Hr glycosyl hydrolase, family 32
	A Glycoside hydrolase, family 32
orf33	Hr glycosyl hydrolase, family 53 - likely arabinogalactan 1,4-beta-galactosidase
	A Arabinogalactan endo-1,4-beta-galactosidase
orf40	Hr short-chain dehydrogenase / reductase family protein
	A Short-chain dehydrogenase / reductase SDR
orf58	Hr peptidase, family M49
	A dipeptidyl-peptidase III
orf62	Hr protein involved in phosphonate metabolism
	A phnA protein
orf69	Hr arylformamidase
orf73	A N-formylkynurenine (aryl-) formamidase
	Hr Holliday junction nuclease RuvC
orf95	A Crossover junction endodeoxyribonuclease RuvC
	Hr RNA pseudouridylate synthase
	A Ribosomal large subunit pseudouridine synthase D

InterPro observation is in fact the same as the Pfam observation (InterPro incorporates Pfam) it uses a slightly different naming scheme. The remaining annotations are based on different BLAST observations which use different wordings.

An estimate of 9% of all created annotations are wrongly considered to be mismatches by the semi-automatic evaluation process when annotations created by MicHanThi and the human annotator were compared. Considering these 9%, MicHanThi reproduced 80% of the annotations that were also created by the human annotator. The remaining 20% are mismatches in which either the computer or the human annotator is correct.

## 7.7 Conclusions

The work described herein offers a reliable foundation for further studies of genomic and metagenomic data. MicHanThi has been successfully applied to more than ten annotation projects handled by the Microbial Genomic Group since the initial annotation of the organism *Gramella forsetii* KT0803 in late 2004 (27; 73; 74).

The possibility to install the software locally offers full flexibility regarding tool selection and adjustment of annotation parameters. Its expandability and focus on the prediction of gene functions makes it possible to easily integrate the software into custom annotation pipelines.

It has been shown, that the reasoning provided by fuzzy logic to evaluate observations in cooperation with clustering of BLAST observations and the implemented handling of InterPro observations is well suited to reproduce human annotations. Compared to the annotations created by the human annotator, more than 70% of the annotations predicted by MicHanThi were syntactically identical and in addition, more than 9% were semantically equivalent. MicHanThi clearly outperforms the *best BLAST observations* approach and has advantages over the RAST system in the class of (*conserved*) *hypothetical proteins*.

Using MicHanThi to create a preliminary set of annotations reduces the manual workload considerably because human annotators only need to manually inspect annotations considered to be uncertain or weak. The reliability values provided for each annotation support this process effectively. Although the expert annotation still outperforms any automatically derived functional assignment, the results show that the overall annotation quality becomes independent of the experience of the annotation team. Especially for multi-genome comparisons high quality automatic annotations systems are a prerequisite to avoid differences due to different annotation approaches and expertise.

## Acknowledgement

The author likes to take the chance to particularly thank Michael Richter, Hanno Teeling, Thierry Lombardot, and Margarete Schüler for providing the high quality annotation of *Gramella forsetii* KT0803. Also, large parts of the algorithm are based on how they annotate genes and genomes.



## Chapter 8

# *Gramella forsetii* KT0803 Paper

### Whole genome analysis of the marine *Bacteroidetes* *Gramella forsetii* reveals adaptations to degradation of polymeric organic matter

M. Bauer<sup>a</sup>, M. Kube<sup>b</sup>, H. Teeling<sup>a</sup>, M. Richter<sup>a</sup>, T. Lombardot<sup>a</sup>, E. Allers<sup>a</sup>,  
C.A. Würdemann<sup>a</sup>, C. Quast<sup>a</sup>, H. Kuhl<sup>b</sup>, F. Knaust<sup>b</sup>, D. Woebken<sup>a</sup>,  
K. Bischof<sup>a</sup>, M. Mussmann<sup>a</sup>, J.V. Choudhuri<sup>c</sup>,  
F. Meyer<sup>c</sup>, R. Reinhardt<sup>b</sup>, R.I. Amann<sup>a</sup>,  
and F.O. Glöckner<sup>a,d</sup>

<sup>a</sup>Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>b</sup>Max Planck Institute for Molecular Genetics, D-14195 Berlin, Germany;

<sup>c</sup>University of Bielefeld, Centre for Biotechnology (CeBiTec), D-33594 Bielefeld, Germany;

<sup>d</sup>International University Bremen, D-28759 Bremen, Germany;

**Journal; Volume (Issue):** Environ. Microbiol.; 8 (12)

**Pages:** 2201-2213

**Month / Year:** October 2006

**DOI:** 10.1111/j.1462-2920.2006.01152.x

#### **Contributions:**

Motivation for the implementation of MicHanThi. First MicHanThi evaluation test case. Validation of all *hypothetical* and *conserved hypothetical proteins*. Syntax checks for the created annotation.

---

**Contents**

---

<b>8.1</b>	<b>Introduction</b>	<b>93</b>
<b>8.2</b>	<b>Results and Discussion</b>	<b>94</b>
8.2.1	General genomes features	94
8.2.2	Hydrolytic capabilities	94
8.2.3	Nutrient sensing	96
8.2.4	Metabolic control during feast and famine conditions	97
8.2.5	Nitrogen-, sulfur- and phosphorus metabolism (summarized in Fig. S4)	98
8.2.6	Optimization of carbon source/nutrient acquisition	99
8.2.7	Potential sensing of light	99
8.2.8	Surface adhesion potential	100
8.2.9	Particle-associated life style	102
<b>8.3</b>	<b>Conclusions</b>	<b>102</b>
<b>8.4</b>	<b>Experimental procedures</b>	<b>103</b>
8.4.1	Sequencing and assembly	103
8.4.2	Gene prediction and annotation of the genome sequence	103
8.4.3	Analysis of the genome architecture	103
8.4.4	Comparative analyses	103
<b>8.5</b>	<b>Acknowledgements</b>	<b>103</b>

---

## Abstract (Summary)

Members of the *Bacteroidetes*, formerly known as the *Cytophaga-Flavobacteria-Bacteroides* (CFB) phylum, are among the major taxa of marine heterotrophic bacterioplankton frequently found on macroscopic organic matter particles (marine snow). In addition, they have been shown to also represent a significant part of free-living microbial assemblages in nutrient-rich microenvironments. Their abundance and distribution pattern in combination with enzymatic activity studies has led to the notion that organisms of this group are specialists for degradation of high molecular weight compounds in both the dissolved and particulate fraction of the marine organic matter pool, implying a major role of *Bacteroidetes* in the marine carbon cycle. Despite their ecological importance, comprehensive molecular data on organisms of this group have been scarce so far. Here we report on the first whole genome analysis of a marine *Bacteroidetes* representative, '*Gramella forsetii*' KT0803. Functional analysis of the predicted proteome disclosed several traits which in joint consideration suggest a clear adaptation of this marine *Bacteroidetes* representative to the degradation of high molecular weight organic matter, such as a substantial suite of genes encoding hydrolytic enzymes, a predicted preference for polymeric carbon sources and a distinct capability for surface adhesion.

## 8.1 Introduction

Vertical mass fluxes in the ocean drive element cycling and are mediated by biogenic particles (109). Main particle sources are phyto- and zooplankton, but particle formation also involves the trapping of organic macromolecules (110). Macroscopic aggregates, known as marine snow (111), are formed in the photic zone and mineralized as they sink deeper to meso- and bathypelagic zones (112). This transports carbon and other nutrients to deeper zones, effectively lowering atmospheric CO<sub>2</sub> in a process known as the 'marine biological pump'. The sequestration of particulate organic matter (POM) therefore has a profound influence on global climate (113).

Organic matter mineralization is mainly catalysed by heterotrophic bacteria (114). Among the major taxa of marine bacterioplankton, members of the *Bacteroidetes*- formerly known as the *Cytophaga-Flavobacteria-Bacteroides* (CFB) phylum – are frequently found enriched on organic matter particles (115; 116), and are increasingly noticed to also dominate free-living microbial assemblages in nutrient-rich microenvironments associated with phytoplankton blooms (117; 118). Studies on both cultivated and uncultivated marine *Bacteroidetes* have shown the ability of members of this group to efficiently consume biopolymers (like protein and the polysaccharide chitin) (119; 120), which make up a significant fraction of the high-molecular-weight dissolved organic matter (DOM) pool in the oceans (121). Biopolymer degradation is considered as the rate limiting step in DOM mineralization by marine microorganisms, and, hence, *Bacteroidetes* are hypothesized to play a key role in this process in the oceans (120).

Genome sequence data would have the potential to aid in the development of more detailed hypotheses on the role of specific members of this important, wide-spread and diverse group of bacteria in biogeochemical element cycling. However, this kind of comprehensive molecular data on marine *Bacteroidetes* has been scarce so far. Recent metagenomic studies report on the distribution and functional analysis of specific *Cytophaga*-like hydrolases in the Sargasso Sea and the Western Arctic Ocean (122), or describe hydrolase-containing genome-fragments of Antarctic marine *Bacteroidetes* (123). To date, no complete genome analysis of a marine *Bacteroidetes* has been published to address the question how far the genetic inventories of members of this phylum reflect general and special capabilities consistent with their anticipated role in the process of organic matter remineralization. Here, we report on the first genome of a marine aerobic heterotrophic representative of the bacterial phylum *Bacteroidetes*, '*Gramella forsetii*' KT0803, which has been isolated from North Sea surface waters during a phytoplankton bloom (124) and is phylogenetically affiliated with the *Flavobacteria*.

## 8.2 Results and Discussion

### 8.2.1 General genomes features

The 3 798 864 bp genome sequence of the coastal bacterioplankter ‘*Gramella forsetii*’ KT0803 was determined by a random whole-genome shotgun approach and predicted to contain 3585 protein coding sequences (CDS, Table 8.1). An overview on the general genome architecture as well as the location of members of prominent paralogous gene families and functionally related gene groups is given in Fig. S1.

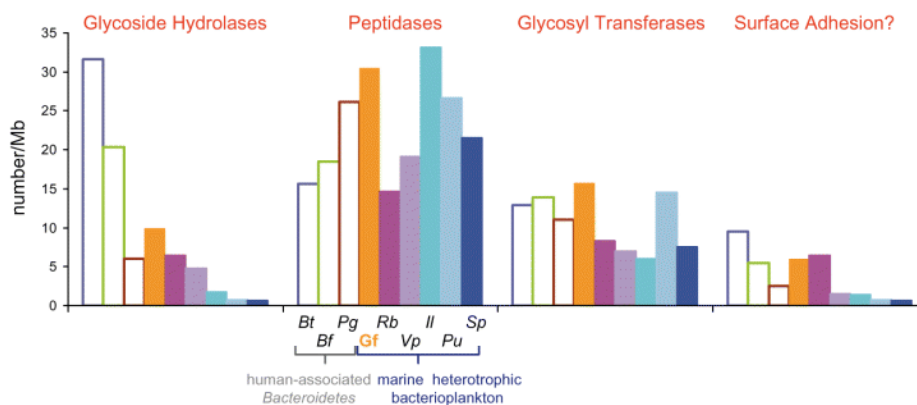
**Table 8.1:** *General features of the ‘Gramella forsetii’ KT0803 genome. a. Covering all proteinogenic amino acids except selenocysteine. b. Gene order 5’ 16S-tRNAIle-tRNAAla-23S-5S 3’. c. tmRNA, RNA component of RNase P, cobalamin riboswitch. CDS, coding sequence; TMH, transmembrane helix; SP, signal peptide.*

Size (bp)	3,798,864	
G + C content (%)	36.6	
Protein-coding genes	3585	
rRNAs <sup>a</sup>	44	
#rRNA operons <sup>b</sup>	3	
structural RNAs <sup>c</sup>	3	
Coding potential (%)	90	
Average CDS length (bp)	954	
CDS with functional assignments (% of total proteins)	1,980	(55.2)
CDS with hints on potential function (% of total proteins)	243	(6.8)
Conserved hypothetical proteins (% of total proteins)	625	(17.4)
Proteins with TMH or SP prediction only (% of total proteins)	280	(7.8)
Hypothetical proteins (% of total proteins)	457	(12.7)

### 8.2.2 Hydrolytic capabilities

Generally, *Bacteroidetes* are considered as efficient utilizers of the biopolymers present in marine high molecular weight (HMW) organic matter, mainly polysaccharides and proteins (119; 120). The analysis of the predicted proteome of ‘*G. forsetii*’ revealed a combination of extensive glycolytic and proteolytic potential which, together with certain membrane transport characteristics (see below), substantiates this long-standing notion. Thus, a comparison of the ‘*G. forsetii*’ genome with those of three human-associated *Bacteroidetes* and five marine heterotrophic bacterioplankters (two alpha-, and two gammaproteobacteria, one planctomycete, Fig. 8.1) shows that (i) ‘*G. forsetii*’ possesses the third-highest number of glycoside hydrolases (GHs) per megabase (Mb) (10.5, total 40) after the two *Bacteroides* spp. which are well-known specialists for polysaccharide degradation, and (ii) it harbours the highest number of peptidases per Mb (30.5, total 116) after the deep sea gammaproteobacterium *Idiomarina loihiensis* which has been proposed to rely mainly on amino acids for carbon and energy supply (125).

Several predicted GHs in ‘*G. forsetii*’ show sequence similarity to enzymes known to hydrolyse polysaccharides that are major constituents of plant and marine algal cell walls (Table S1). Arabinose polymers seem to be a particularly well-suited carbon source for ‘*G. forsetii*’: five out of 40 GH are putatively acting on compounds like arabinans, arabinogalactans and arabinoxylans. In contrast to the observed effective degradation of chitin by coastal populations of *Bacteroidetes* (119), and the widespread uptake of the chitin monomer N-acetyl-D-glucosamine by coastal *Bacteroidetes* and pelagic marine bacteria (119; 126), ‘*G. forsetii*’ seems to lack the enzymatic equipment for degradation of this abundant polymer and utilization of its oligo/monomeric hydrolysis prod-



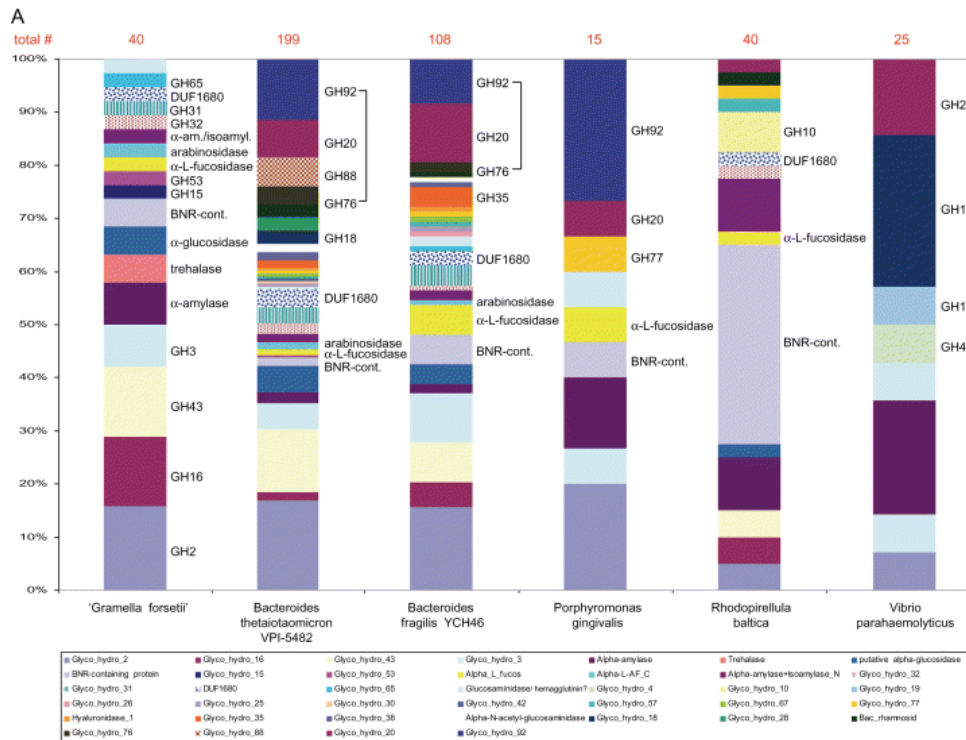
**Figure 8.1:** Comparison of hydrolytic capabilities and adhesion potential of marine heterotrophic bacterioplankton and human-associated members of the Bacteroidetes. Bt, *Bacteroides thetaiotaomicron*; Bf, *Bacteroides fragilis* YCH46; Pg, *Porphyromonas gingivalis*; Gf, ‘*Gramella forsetii*’; Rb, *Rhodopirellula baltica*; Vp, *Vibrio parahaemolyticus*; Il, *Idiomarina loihiensis*; Pu, *Pelagibacter ubique* HTCC1062; Sp, *Silicibacter pomeroyi* DSS-3.

ucts. Regarding proteolytic activities, it is apparent from comparative bacterioplankton peptidase profiles (Fig. 8.2B) that family M14 carboxypeptidases and family S9 prolyl-oligopeptidases play a more prominent role in the peptidase set of ‘*G. forsetii*’ than trypsin-like proteins and proteases of the families M19, M20, M24 which are represented markedly higher in the proteobacterial and/or planctomycetal peptidase sets (127).

The majority of hydrolytic enzyme sequences in ‘*G. forsetii*’ exhibits a predicted signal peptide, and peptidase sequences frequently also contain predicted transmembrane domains (Table S1, boxed categories in Fig. 8.2B), which implies that these proteins may be involved in the processing of extracellular biopoly and/or oligomers. Organism specific profiles for hydrolytic enzymes such as those becoming apparent in Fig. 8.2A and B might be indicative of specific substrate niches colonized by different members of marine bacterioplankton, reflecting to some extent the composition of the nutrient pool in different ocean provinces and marine microenvironments. On the other hand, these organismal differences could explain the presence or absence of distinct hydrolytic activities in certain habitats depending on the microbial community colonizing them (128).

Polymeric nutrient binding by outer membrane complexes. Intriguingly, more than half of the GH and also one peptidase (Table S1) are encoded in the direct vicinity of (among others) two characteristic outer membrane proteins that exhibit similarity to SusC and SusD from *Bacteroides thetaiotaomicron*, which in this organism function as polysaccharide binding entities in a multicomponent outer membrane starch utilization system [Sus (129)]. SusC belongs to the TonB-dependent outer membrane receptor family, one of the most extended paralogue families (40 members) in the ‘*G. forsetii*’ genome, in sharp contrast to the genome of two other marine coastal heterotrophic bacterioplankton, *Pelagibacter ubique* and *Silicibacter pomeroyi* which encode no members of this family. Moreover, the genetic context of the 14 *susCD*-like operons in ‘*G. forsetii*’ (Table S1) frequently encodes exported proteins with PKD-domains (presumably involved in extracellular protein–protein interaction). These findings suggest that ‘*G. forsetii*’ employs a similar strategy as *B. thetaiotaomicron* for binding and degradation of polymers (carbohydrate and putatively also protein) by cell surface complexes.

Although there are instances of solute specific transport systems in ‘*G. forsetii*’, such as 18 major facilitator superfamily (MFS) secondary active transporters (mainly for sugar mono/oligomers but also for oligopeptides), and 28 (7.4/Mb) ABC-type primary

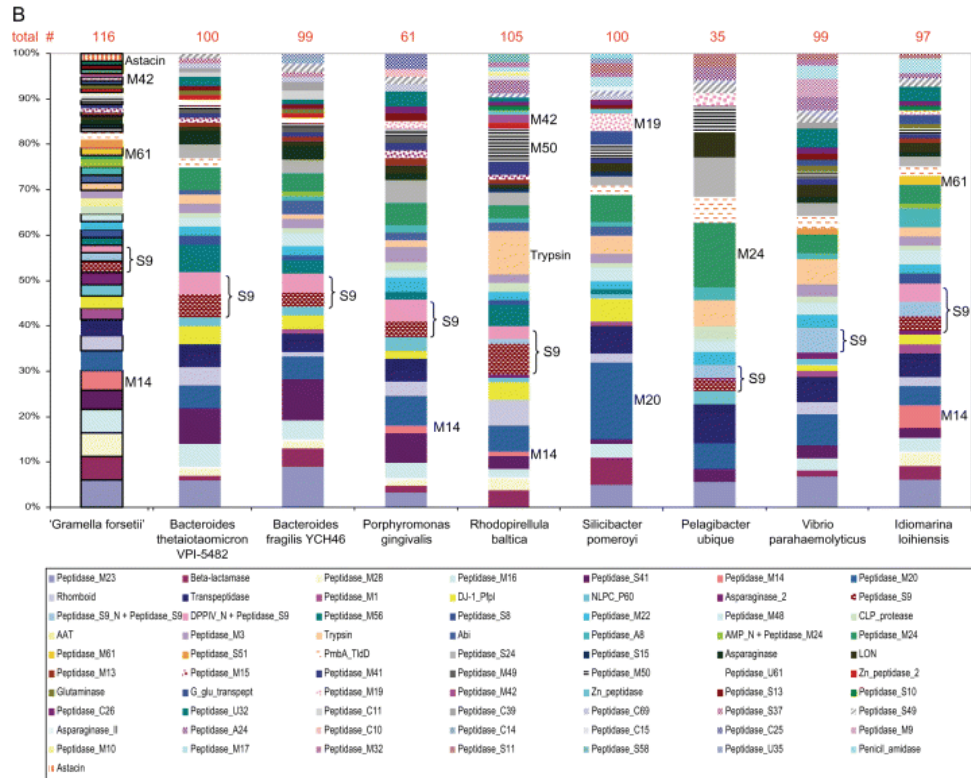


**Figure 8.2:** Comparison of gene family profiles of marine heterotrophic bacterioplankton and human-associated members of the Bacteroidetes. The contribution of special families to an organism's total set is shown for: A. Glycoside hydrolases. B. Peptidases (boxed categories: more than 50% of family members in '*G. forsetii*' are predicted with a signal peptide and/or one or more transmembrane helices). C. Glycosyl transferases.

active transporters, '*G. forsetii*' was found to completely lack periplasmic solute binding proteins of certain families (Pfam families SBP\_bac\_1, \_3, \_5, \_7, \_9; Table S2). This constitutes a glaring contrast to *S. pomeroyi* (130) with its 59 (13/Mb) periplasmic solute binding proteins and 102 (22/Mb) ABC-type transporters, and is also distinctly different from what is found in the oligotrophic bacterioplankton *P. ubique* (131). Normalized to genome-size, *P. ubique* possesses a more than twofold higher number of ABC-type transport systems than '*G. forsetii*', and encodes per Mb 6 periplasmic solute binding proteins (Table S2). Although an unusual occurrence of multiple paralogues in some transporter families has been noted for *alphaproteobacteria* (132; 133), the distinct variations in major transport protein families between *P. ubique*, *S. pomeroyi* and '*G. forsetii*' may well reflect profound differences in nutrient utilization strategies: while the former are well equipped with highly specific and affine extracytoplasmic solute binding receptors to efficiently use a broad variety of monomers even at low concentrations in an oligotrophic ocean, the latter seems more adapted to rely on a set of polymeric nutrients, for which outer membrane receptors may confer specificity, and which provide high concentrations of oligo/monomers after initial degradation.

### 8.2.3 Nutrient sensing

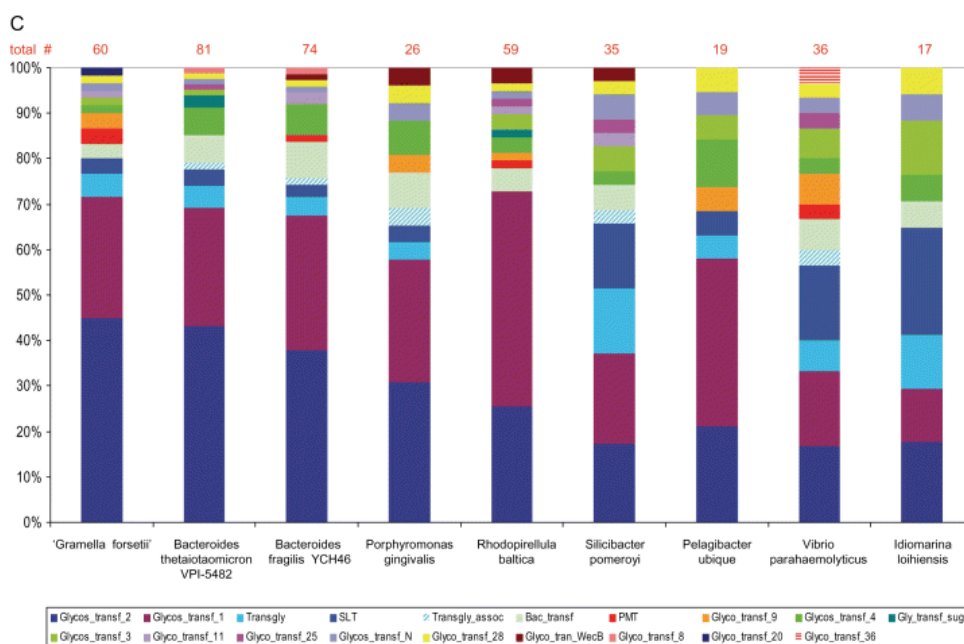
Judged by the adjacent genome localization of the respective genes, several polysaccharide degradation activities in '*G. forsetii*' seem to be regulated by environmental stimuli via conspicuous hybrid-sensor/response regulator signal transduction systems (Tables S2 and S4; Fig. S2) reminiscent in domain composition (Table S3) of the so-called



one-component-systems in *B. thetaiotaomicron* (134). The advantage of such all-in-one signal transduction systems has been attributed to a gain in specificity (reduced risk of cross-talk between sensing systems) under conditions where the concomitant loss of sensitivity (reduced amplification of nutrient signal) is irrelevant, e.g. when high concentrations of a specific nutrient prevail (134). Thus, '*G. forsetii*' likely experiences phases of high polysaccharide nutrient supply of confined chemical nature, perhaps during bloom and senescence phases of a specific phytoplankton population. This kind of specialization for a defined set of polymeric nutrient compounds would be consistent with the results of seawater mesocosm studies in which a taxonomic linkage between blooming phytoplankton and associated bacterial populations has been demonstrated (135). An intriguing alternative way of sensing macromolecules and initiating an appropriate molecular response without need for prior uptake/degradation of the macromolecule could be via so-called TonB-dependent transducers, TonB-dependent outer membrane receptors interacting through an N-terminal extension with a FecRI-like ECF-sigma/anti-sigma system (136). In fact, '*G. forsetii*' harbours three potential systems of this kind (encoded by orf1240–1242, orf1247–1249 and orf1717–1719), of which the latter could be relevant with respect to glycolytic activities of '*G. forsetii*' on marine polysaccharides because it is encoded next to a putative kappa-carrageenase (orf1712, Table S1).

#### 8.2.4 Metabolic control during feast and famine conditions

Marine *Bacteroidetes* seem to be capable of rising from low cell densities in nutrient-poor blue waters (22) to high abundance when nutritional conditions improve (115; 137). It has been shown that members of the *Bacteroidetes* seem to be particularly responsive to an increase in concentration of HMW DOM (138), in contrast to the oligotrophic bacterium *P. ubique*, which does not show an increase of growth rate upon nutrient addition (131). The presence of three rRNA operons in the genome of '*G. forsetii*' is consistent with the potential to respond with rapid protein synthesis to an increase in resource availability (139). Under laboratory conditions, the organism shows a relatively short doubling time



of 250 min in complex medium. *'Gramella forsetii'* has been isolated during a North Sea late-summer phytoplankton bloom (124), in a season during which it contributed with  $6\% \pm 2\%$  to total picoplankton cell counts together with related *Bacteroidetes*. Its genome reveals the potential for metabolic control during 'feast' conditions, as they occur in the course of an algal bloom, as well as for survival of 'famine' conditions, prevailing between algal-blooms. Generally, the organism appears to have a flexible and adjustable metabolism with isoforms of key enzymes encoded in different genetic contexts (Fig. S3). This suggests that the metabolic flux is extensively regulated dependent on the organism's physiological status due to growth conditions and/or developmental phase. In particular, the genome encodes the methylglyoxal synthesis/detoxification route (Fig. S3A), suitable to prevent loss of control over carbon flux during rapid environmental changes from 'famine' to 'feast' conditions by withdrawing triosephosphate from the glycolytic pathway and thereby counteracting the accumulation of toxic phosphorylated intermediates (140). A life style between alternating phases of excess and deprived nutrient supply is further suggested by the organism's enzymatic potential to synthesize storage compounds for carbon, nitrogen (glycogen synthase Orf3500, cyanophycin synthetase Orf735), and phosphorus (polyphosphate kinase, Orf1608, Orf1817), and to remobilize nutrients from this store (Fig. S4).

Like *S. pomeroyi*, the heterotrophic organism *'G. forsetii'* harbours in its genome two operons encoding aerobic carbon monoxide hydrogenase complexes (*coxSML*: orf435–437, *coxSL*: orf438–439). A potential ability of *'G. forsetii'* to oxidize carbon monoxide for energy gain would be consistent with the recent identification of a member of the *Bacteroidetes* (class *Sphingobacteria*) among the microbial community capable of carbon monoxide oxidation in a coastal marine surface water environment (141). More generally, it would suggest a wider occurrence of *Silicibacter*-like lithoheterotrophy for supplementary energy generation in the marine realm (130).

### 8.2.5 Nitrogen-, sulfur- and phosphorus metabolism (summarized in Fig. S4)

According to the predicted enzymatic instrumentation of *'G. forsetii'*, neither nitrate nor nitrite is assimilated. The organism seems to rely solely on reduced nitrogen sources (ammonium, and organically bound nitrogen in amino acids, but not urea, Fig. S4A).



In this respect, '*G. forsetii*' is similar to *P. ubique* and *S. pomeroyi* which also lack the genes for nitrate/nitrite assimilation, albeit both of them seem capable of utilizing urea. Whereas '*G. forsetii*' possesses only one ammonium transporter gene, the two *alphaproteobacterial* genomes encode four genes each. This functional redundancy might reflect a greater dependence on ammonium as a nitrogen source and concomitantly the need to deal with low concentrations of this scarce nutrient in the oligotrophic ocean. In contrast, '*G. forsetii*' might encounter rich sources of dissolved organic nitrogen-like protein, and utilize it by its proposed protein degradation capabilities to meet its nitrogen demand.

Dissimilatory use of nitrogen compounds in '*G. forsetii*' is apparently restricted to nitrous oxide (Fig. S5). Intriguingly, the nitrous oxide reductase gene (*nosZ*, orf1398) resides in a region of the genome that is dedicated to alleviate the effects of severe oxygen limitation (Fig. S5B), harbouring a potential oxygen-sensitive regulator system analogous to the FixLJ two-component system (Fig. S5C, orf1406 / orf1407) and two instances of an extremely oxygen-affine terminal respiratory oxidase type [*cbb3*, Fig. S5B (142; 143)]. The region also encodes a potential regulator similar to *Pseudomonas stutzeri* DnrN, which is part of the nitric oxide-dependent regulation (DnrD) operon, where DnrD is the global NO-dependent regulator of nitrite denitrification gene expression. In this system, NO functions (via DnrD) as a coinducer also for nitrous oxide reductase expression (144). Possibly, '*G. forsetii*' can use nitrous oxide produced by anaerobic nitric oxide detoxification systems of other members of the microbial community (145) as a terminal electron acceptor to gain energy even at extremely reduced oxygen concentrations.

The genome of '*G. forsetii*' features a further remarkable detail, which is possibly linked to nitrogen metabolism: a gene encoding a potential deoxyhypusine synthase (DHS, orf3578; Fig. S5A). Genes encoding proteins similar to this typically eukaryal and archaeal post-translational modification enzyme have been detected in only very few instances in *Bacteria*, and their function in this domain is enigmatic (146). Considering its genetic context (orf3580, orf3579), the DHS-like protein in '*G. forsetii*' might participate in the biosynthesis of the polyamine homospermidine from arginine: homospermidine has been found to be a typical polyamine in members of the *Bacteroidetes* (147), yet a canonical homospermidine synthase could not be detected in the '*G. forsetii*' genome. Interestingly, a secondary metabolism-specific homospermidine synthase has been shown to originate from DHS in plants (148).

The sulfate assimilation pathway in '*G. forsetii*' is interesting with respect to the apparent lack of APS-kinase, suggesting that assimilatory sulfate reduction occurs at the APS-level by an APS-reductase (149). The genome of '*G. forsetii*' contains two sulfatase genes (orf356, orf1682), which are located in close proximity to glycosylidase genes (Fig. S2, Table S1) suggesting that certain sulfated polysaccharides may be a substrate for '*G. forsetii*'.

### 8.2.6 Optimization of carbon source/nutrient acquisition

Inherently, the proposed mode of carbon source forging by attachment and concomitant hydrolysis and uptake of the break-down products (Sus-like outer membrane complexes) avoids release of hydrolytic enzymes to the surrounding environment and minimizes the loss of utilizable substrate to other members of the microbial community.

### 8.2.7 Potential sensing of light

Intriguingly, the '*G. forsetii*' genome harbours two instances of potential metabolic regulation by light: (i) a blue light and/or redox status sensor protein [Orf374 (150)] encoded next to a metabolic key enzyme catalysing the first step of the non-oxidative part of the pentose phosphate pathway (ribulose-5-phosphate-epimerase, Orf373, Fig. S3B), and (ii) a phytochrome-like photoreceptor (Orf1940), which is rather unusual for heterotrophic bacteria (151), as part of a two-component signal transduction system. Although the

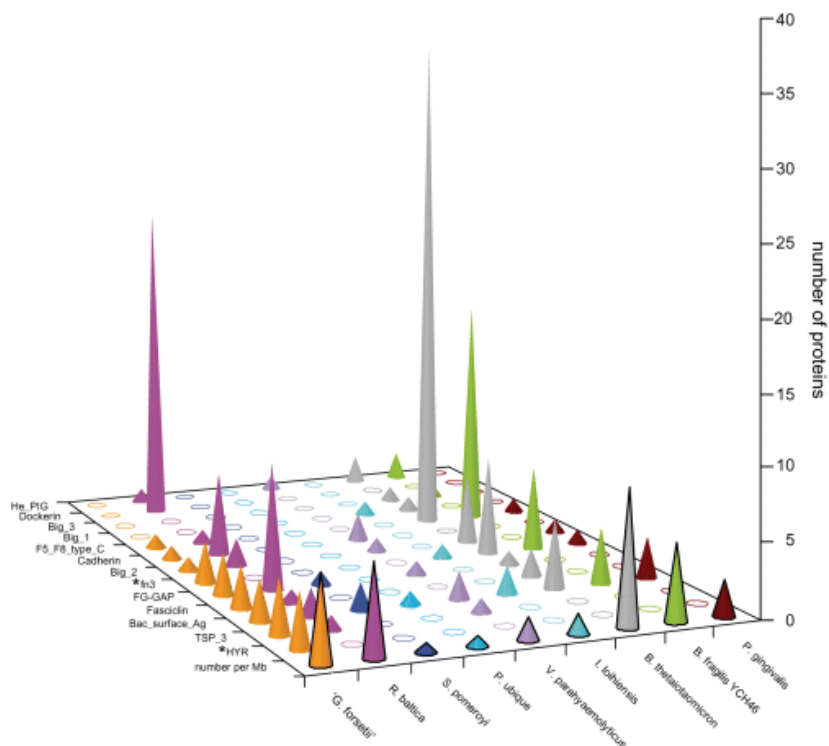
type of metabolic response that might be elicited by light stimuli in '*G. forsetii*' is not clear, it is tempting to speculate about such a regulation because nutrient concentrations are higher in the euphotic zone of aquatic systems, marine *Bacteroidetes* are frequently observed associated with phytoplankton populations, and a potential metabolic regulation in response to light seems principally advantageous to an organism feeding on either exudates or detritus of light-dependent photosynthetic organisms like marine algae. The bacteriophytochrome in '*G. forsetii*' is encoded next to one of two sensor histidine kinases (Orf1941 and Orf1669) whose sequences, quite unusually, contain domains of chemotactic adaptation proteins (methyltransferase CheR and methylesterase CheB, Table S3). Otherwise, typical chemotaxis genes could not be detected in the genome, hence '*G. forsetii*' is either incapable of chemotactic response or this phenomenon is brought about by a novel mechanism involving an as yet unknown set of proteins. Although '*G. forsetii*' possesses all but one of the so far known gliding motility genes (Fig. S6; see (152)), motility has not been observed in this organism (S. Verburg and B. Tindall, pers. comm.). The presence of genes encoding sensor/regulator systems in the genetic context of several potential gliding motility genes (Fig. S6) suggests, however, that a potential onset of motility requires specific stimuli.

### 8.2.8 Surface adhesion potential

Several environmental studies have found marine *Bacteroidetes* primarily as members of particle-associated microbial communities (115; 116). Consistent with this notion, the genome of '*G. forsetii*' harbours extensive potential for surface adhesion. Thus, after *B. thetaiotaomicron* and *Rhodopirellula baltica*, '*G. forsetii*' possesses the highest number of predicted proteins with domains that have been implicated in cell-cell and cell-surface interaction (Figs 8.1 and 8.2). The characteristic repeat structures of some of these proteins are known to bind calcium ions [e.g. thrombospondin type 3 repeat (TSP\_3), cadherin]. Cell surface structures formed by these proteins could be a means to adhere to algal surface mucilage and thus to colonize living phytoplankton cells. Recently, it has been suggested that the formation of marine snow particles is strongly triggered by (algal) polysaccharide aggregation, which, in turn, is catalysed by metal cation bridging (e.g. calcium) (110). Although speculative at this time, it is tempting to assume that certain surface structures of '*G. forsetii*' could mediate interaction with acidic organic macromolecules such as algal exudates, thereby facilitating an attachment to nascent nutrient-rich marine snow particles.

As a further asset with respect to surface attachment, '*G. forsetii*' encodes 60 glycosyl transferases, the majority of which is likely to participate in biosynthesis of cell wall components including extracellular polysaccharide structures known to mediate cell/cell as well as cell/surface adhesion (155). Interestingly, *P. ubique*, although known to grow as unattached cells suspended in the water column, shows also a relatively high number of glycosyl transferase genes per Mb (Fig. 8.1) A possible reason for this apparently 'over-represented' functional group could be that the generation of a functional cell wall needs a certain number of glycosyl transferases that even an organism with extremely small genome [1.31 Mb (131)] has to afford in order to be protected in a frequently hostile environment. Also, by their physicochemical properties membrane polysaccharides might aid in the binding of scarce trace metals, an additional asset in the oligotrophic habitat of *P. ubique*.

Although '*G. forsetii*' possesses no quorum sensing system which is used by other bacteria to time the physiological switch from the free-living to the attached mode (156), there are indications that exopolymer biosynthesis and restructuring are triggered by environmental signals in '*G. forsetii*' and may be connected to an incising developmental switch possibly represented by the transition from a free-living to an adherent life-style, e.g. in biofilms on nutrient-rich particles: exopolymer biosynthesis genes in '*G. forsetii*' are organized in several operons (Fig. S1) of varying genetic contexts, some of which are colocalized with systems of environmental signal transduction suggesting differential



**Figure 8.3:** Comparison of abundance and types of proteins potentially mediating surface adhesion in marine heterotrophic bacterioplankton and human-associated members of the Bacteroidetes. Domain explanations are available at <http://www.sanger.ac.uk/Software/Pfam/tsearch.shtml>. \*Note: (i) Fibronectin type 3 homology domains (*fn3*) are also found in bacterial extracellular glycoside hydrolases. Recently, it has been shown that *fn3* domains in a bacterial extracellular glycoside hydrolases function in promoting the hydrolysis of the polysaccharide by modifying its surface (153). In ‘*G. forsetii*’, *fn3* domains are present in two hydrolytic enzymes [Orf363 (*GH\_43*), Orf1026 (*peptidase\_M28*)] but also in a non-enzymatic exported protein (Orf2394). (ii) One of the hyalin-repeat containing proteins (Orf370) exhibits also a cell-surface attachment module similar to the C-terminal domain V of the modular xylanase *Xyn10A* from *Rhodothermus marinus* (154).

expression with potential modification of the cell surface structure upon reception of certain external stimuli. Externally triggered surface modification as important process in the life style of the organism is further underpinned by (i) the presence of a protein (Orf148) containing a domain that is predicted to sense stimuli that are specific for the developmental program of an organism (CHASE, (157)) in the vicinity of an exopolymer biosynthesis protein, and (ii) the prominent representation of the two component system effector domain LytTR (nine out of 29 known output domains, Table S3) in the set of signal transduction systems. This domain has primarily been found in transcriptional regulators that are involved in biosynthesis of extracellular polysaccharides, fimbriation, expression of exoproteins (including toxins), and quorum sensing (158). By virtue of its potential to synthesize exopolymers, '*G. forsetii*' might also play a role in particle aggregation and/or the stabilization of existing aggregates, thus promoting carbon sequestration rather than accelerating its remineralization (159).

### 8.2.9 Particle-associated life style

Two of the several systems conferring adaptation to fluctuating and adverse environmental conditions (Table S5) in the genome of '*G. forsetii*' may have relevance to a potential particle-associated life-style: (i) a flexible respiratory chain (Fig. S5) with several terminal oxidases [cytc(CuA)-, cytd-, cbb3-type] apt to function under different regimes of ambient oxygen concentration – advantageous because marine snow particles can transiently become micro to anoxic due to high respiratory activity and limited oxygen diffusion (160); (ii) a large number (Table S1) of primary (7 cation-translocating P-type ATPases) and secondary (8 of 11 RND-type export systems) active transport systems for the detoxification of heavy metal ions – beneficial in a polluted habitat, but potentially also during particle attachment because organic macromolecules trapped in and triggering formation of marine snow particles are able to scavenge trace metal ions (110; 161), possibly demanding efficient means to ensure metal homeostasis from an organism growing attached to these particles.

## 8.3 Conclusions

The genome analysis of '*G. forsetii*' KT0803 presented here allows a first glimpse into the genetic potential of marine members of the *Bacteroidetes*. With its substantial suite of genes encoding glycolytic and proteolytic enzymes, a predicted preference for polymeric carbon sources and a distinct capability for surface adhesion, the organism seems capable of a life-style that would be consistent with a repeatedly formulated hypothesis on the ecological role of marine *Bacteroidetes* (120) which is to initiate remineralization of HMW organic matter either from the particulate or the dissolved organic carbon pool. However, the predicted inability of '*G. forsetii*' to utilize the abundant polysaccharide chitin and/or its breakdown products is in contrast to what has been reported for the *Bacteroidetes* community in general. Also, none of the GH families found to be most abundant in the Sargasso Sea metagenomic data set (122), is represented in '*G. forsetii*'. On the other hand, several GH genes of '*G. forsetii*' do have similar counterparts in the Sargasso Sea (M.B., in preparation). These findings suggest that degradative capabilities of specific *Bacteroidetes* are most likely focused on distinct polymer/oligomer fractions and that there may be a more or less sharp biopolymer substrate niche speciation within the *Bacteroidetes* community. The present analysis is the first to report on organism specific profiles of predicted biopolymer degrading activities within the marine *Bacteroidetes*. This opens up the field for comparative studies to address the important question, which groups of the heterotrophic bacterioplankton are responsible for the degradation of which fraction of HMW organic matter in aquatic systems, and to which extent biopolymer substrate spectra overlap between species. It will be very interesting to see how these patterns of predicted hydrolytic activities differentiate once additional

planktonic genomes of the *Bacteroidetes* become available.

## 8.4 Experimental procedures

### 8.4.1 Sequencing and assembly

A whole genome shotgun sequencing approach was chosen for the analysis of the '*Gramella forsetii*' KT0803 genome. Three plasmid libraries (average insert sizes of 1.2 kb, 2.5 kb and 3.4 kb) were generated. DNA preparation, sequencing and assembly was performed as described earlier (162). Overall, 64 653 high quality sequence reads were generated during the phases of shotgun sequencing, gap closure and sequence/assembly verification. Resequencing and primer-walking in combination with fosmid internal sequence analysis was used as a strategy for improving regions of weak quality and for gap closure. Sequence information could be assembled into a single contig of 3 798 864 bp, with 9.8-fold sequence coverage and a sequence quality of less than 1 error in 100 000 bp.

### 8.4.2 Gene prediction and annotation of the genome sequence

For the final annotation, potential protein coding genes (open reading frames, orfs) were identified using the in house gene prediction software mORFind (Waldmann and Teeling, unpublished). tRNA genes were identified using tRNAScan-SE (163). Ribosomal RNA genes were detected by standard similarity searches [blast, (7)] against public nucleotide databases, and additional structural RNA species were identified by similarity searches of intergenic regions against the Rfam database (164). Annotation of the genome sequence was performed with the GenDB v2.0 system (9) collecting for each predicted ORF observations from similarity searches against sequence databases (nr, nt, SWISS-PROT) and protein family databases (Pfam, Prosite, InterPro, COG), and from predictive signal peptide- [SignalP v2.0 (165)] and transmembrane helix-analysis [TMHMM v2.0 (63)]. Predicted protein sequences (Orfs) were automatically annotated using a fuzzy logic-based approach including evaluation and integration of analysis tool results (Quast, 2006). (10). In addition, this automatic annotation was manually checked and refined for each Orf.

### 8.4.3 Analysis of the genome architecture

Cumulative GC-skew  $[(G - C)/(G + C)]$  and genome-wide fluctuations of GC-content and tetranucleotide-frequency were calculated with custom made Perl scripts using a window averaging approach.

### 8.4.4 Comparative analyses

Paralogous protein families within the '*G. forsetii*' genome were identified by blast all-against-all similarity comparisons at a significance cut-off level (E-value) of  $10^{-4}$ . Intergenomic protein family comparisons are based on matches to specific Hidden Markov Models (HMMs) in the Pfam database version 18.0, with an E-value cut-off of  $10^{-4}$ .

## 8.5 Acknowledgements

We greatly appreciate the substantial contribution of J. Waldmann (University of Münster, Germany) to the generation of the gene prediction software mORFind, as well as the genome annotation support by I. Kostadinov and M. James (International University Bremen). J. Wulf (MPI Bremen) is acknowledged for cell culture and DNA preparation. Financial support for this work came from the German Max Planck Society and the European Union (Network of Excellence Marine Genomics Europe).



## Chapter 9

# *Congregibacter litoralis* KT71 Paper

### Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis.

B.M. Fuchs<sup>a</sup>, S. Spring<sup>b</sup>, H. Teeling<sup>a</sup>, C. Quast<sup>a</sup>, J. Wulf<sup>a</sup>,  
M. Schattenhofer<sup>a</sup>, Y. Shi<sup>a</sup>, S. Ferreira<sup>c</sup>, J. Johnson<sup>c</sup>, F.O. Glöckner<sup>a,d</sup>,  
and R. Amann<sup>a</sup>

<sup>a</sup>Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>b</sup>Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ), D-38124 Braunschweig,  
Germany;

<sup>c</sup>J. Craig Venter Institute, MD 20850 Rockvill, U.S.A.;

<sup>d</sup>International University Bremen, D-28759 Bremen, Germany;

**Journal; Volume (Number):** Proc. Natl. Acad. Sci. U.S.A.; 104 (8)

**Pages:** 2891-2896

**Month / Year:** February 2007

**DOI:** 10.1073/pnas.0608046104

#### **Contributions:**

Second MicHanThi evaluation test case and base for realised enhancements. Support during the annotation of the genome (automatic annotation). Syntax checks for the created annotation.

---

**Contents**

---

<b>9.1</b>	<b>Introduction</b>	<b>107</b>
<b>9.2</b>	<b>Results and Discussion</b>	<b>109</b>
9.2.1	Structure and Phylogenetic Analysis of the Photosynthesis (PS) Operon.	109
9.2.2	Pigment Analysis	110
9.2.3	Photoautotrophy vs. Photoheterotrophy.	110
9.2.4	Putative Regulation of PS.	111
9.2.5	Microaerophily.	111
9.2.6	Substrate Spectrum.	111
9.2.7	Storage Compounds.	112
9.2.8	Formation of Aggregates and Polysaccharide Production.	113
9.2.9	Sulfur Metabolism.	113
9.2.10	<i>C. litoralis</i> : A Typical Shelf Bacterium?	113
9.2.11	Significance of the NOR5/OM60 Clade Represented by KT71.	114
<b>9.3</b>	<b>Materials and Methods</b>	<b>115</b>
9.3.1	Sequencing and Assembly.	115
9.3.2	Analysis of the Genome Architecture.	115
9.3.3	Gene Prediction and Annotation of the Genome Sequence.	115
9.3.4	Comparative Sequence Analyses.	115
9.3.5	Physiological Tests.	116
9.3.6	HPLC Analysis of Photosynthetic Pigments.	116
<b>9.4</b>	<b>Acknowledgments</b>	<b>116</b>

---



## Abstract

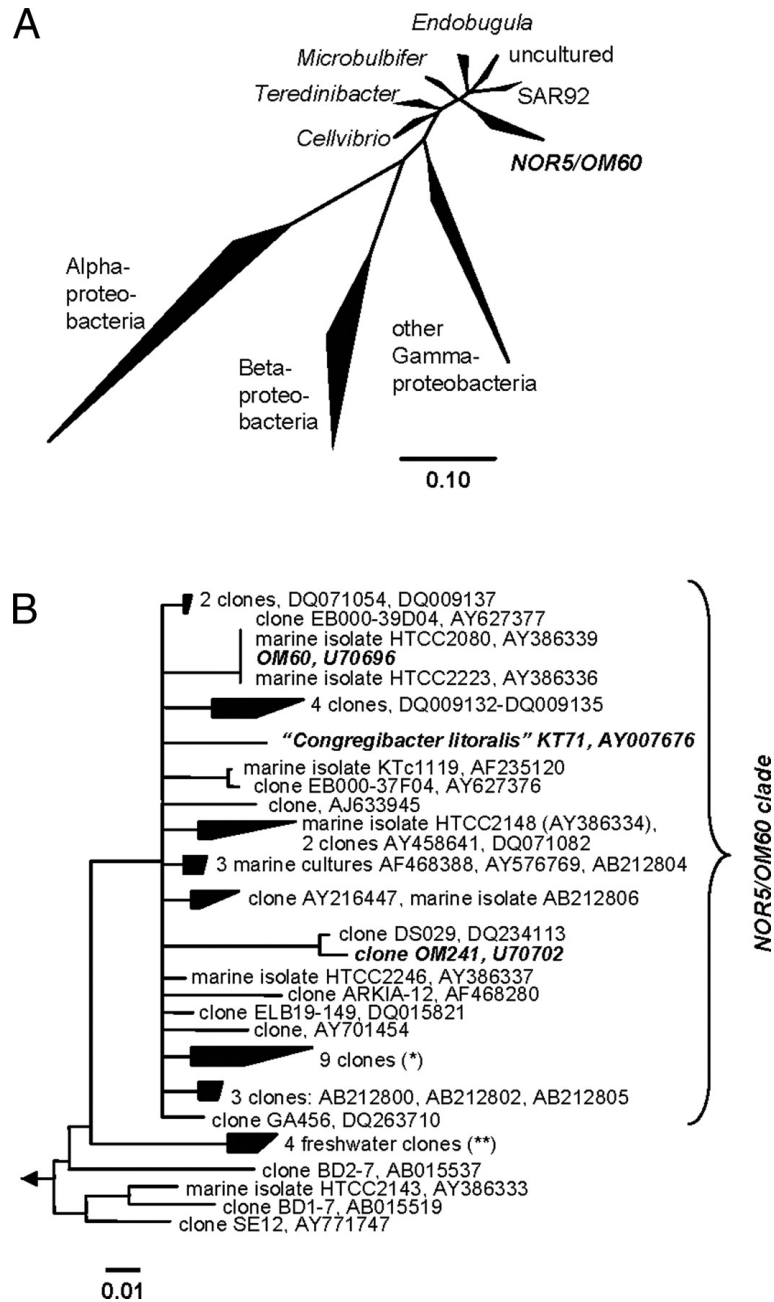
Members of the gammaproteobacterial clade NOR5/OM60 regularly form an abundant part, up to 11%, of the bacterioplankton community in coastal systems during the summer months. Here, we report the nearly complete genome sequence of one cultured representative, *Congregibacter litoralis* strain KT71, isolated from North Sea surface water. Unexpectedly, a complete photosynthesis superoperon, including genes for accessory pigments, was discovered. It has a high sequence similarity to BAC clones from Monterey Bay [Beja O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, et al. (2002) *Nature* 415:630–633], which also share a nearly identical gene arrangement. Although cultures of KT71 show no obvious pigmentation, bacteriochlorophyll a and spirilloxanthin-like carotenoids could be detected by HPLC analysis in cell extracts. The presence of two potential BLUF (blue light using flavin adenine dinucleotide sensors), one of which was found adjacent to the photosynthesis operon in the genome, indicates a light- and redox-dependent regulation of gene expression. Like other aerobic anoxygenic phototrophs (AAnPs), KT71 is able to grow neither anaerobically nor photoautotrophically. Cultivation experiments and genomic evidence show that KT71 needs organic substrates like carboxylic acids, oligopeptides, or fatty acids for growth. The strain grows optimally under microaerobic conditions and actively places itself in a zone of  $\approx 10\%$  oxygen saturation. The genome analysis of *C. litoralis* strain KT71 identifies the gammaproteobacterial marine AAnPs, postulated based on BAC sequences, as members of the NOR5/OM60 clade. KT71 enables future experiments investigating the importance of this group of gammaproteobacterial AAnPs in coastal environments.

## 9.1 Introduction

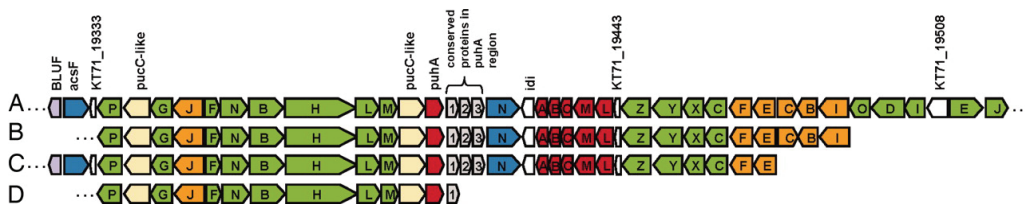
In 1999, Eilers et al. (124) isolated a bacterial strain designated KT71 from a surface water sample taken near the North Sea island Helgoland, by direct plating on complex low-nutrient media. Phylogenetic analysis showed that KT71 was the first cultured representative of a cosmopolitan gammaproteobacterial lineage, which we in the following refer to as the NOR5/OM60 clade (Fig. 9.1). The first indication for this clade dates back to 1997, when Rappe et al. retrieved two 16S rRNA clones, OM60 and OM241, from the continental shelf off Cape Hatteras, NC (166). In the following years, many sequences have been retrieved that were related to the clone OM60 (e.g., refs (167; 168; 169; 170; 171; 172)). By the end of 2005, >180 partial and full length 16S rRNA sequences available within the public databases were related to KT71 and OM60.

KT71 is a Gram-negative, pleomorphic, strictly aerobic, and motile bacterium. It is of an average size of  $2 \times 0.5 \mu\text{m}$ , has a generation time of 4.5 h, and often grows in flocs. Based on this conspicuous trait and the site of isolation, the name *Congregibacter litoralis* has been proposed. A full taxonomic description of strain KT71 is currently ongoing (B.M.F., S.S., and R.A., unpublished work). Several strains belonging to the NOR5/OM60 clade were isolated off the coast of Oregon, in sterilized seawater, using a high-throughput dilution-to-extinction technique (174; 175). Meanwhile, representatives of the NOR5/OM60 clade were also isolated from Arctic sea ice (176) and coastal sediments (177; 178).

FISH with rRNA-targeted oligonucleotide probes for NOR5/OM60 confirmed this clade as an abundant component of the bacterioplankton community in the North Sea around the island Helgoland (124). By the end of July 1998, up to 8% of the total bacterioplankton community comprised members of the NOR5/OM60 clade (124). A second peak of NOR5/OM60 cells was visible in mid-June (6%). However, NOR5/OM60 was not detected by FISH during the winter months, October to March, suggesting a marked seasonality. The fraction of DNA-synthesizing NOR5/OM60 cells seems to be quite variable. Active DNA synthesis could be detected in August but not in May 2002, even though NOR5/OM60 was present in high numbers in both samples (6% and 11% of total bacterioplankton cell, respectively) (179).



**Figure 9.1:** Phylogenetic affiliation of KT71. (A) Parsimony tree of the NOR5/OM60 clade including representative neighboring clades within the Gammaproteobacteria. (B) Consensus tree of the NOR5/OM60 clade reconstructed with 86 almost-full-length sequences (>1,350 nt). All treeing methods and filters resolved a stable branching order for the NOR5/OM60 clade within the Gammaproteobacteria. Within the NOR5/OM60 clade, the branching order could not be unambiguously resolved based on the currently available dataset, which is indicated by a multifurcation (173). (\*), AY212565, DQ015838, DQ015860, DQ015829, DQ015807, DQ015840, AY135664, AY135666, and AY135673; (\*\*), AY212617, AY212664, AY693815, and AY212676.



**Figure 9.2:** Comparison of PS operons. (A) KT71. (B–D) BAC clones EBAC65D09 (B), EBAC29C02 (C), and EBAC69B03 (D). Green, *bch* genes; red, *puf* gene; orange, *crt* genes; blue, hem genes; 1, similar to 23.7-kDa protein (KT71\_19398); 2, similar to 17.4-kDa protein (KT71\_19403); 3, similar to 30.4-kDa protein (KT71\_19408). Dots indicate the presence of genes not belonging to the PS operon.

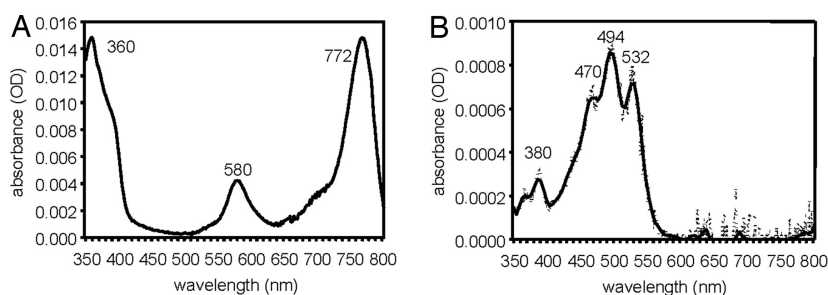
In 2004, KT71 was selected for whole-genome sequencing as part of the Gordon and Betty Moore Foundation (GBMF) Marine Microbiology Initiative. Here, we present data derived from the analysis of the genome of strain KT71 and from ecophysiological experiments addressing some of the predictions derived from genome annotation.

## 9.2 Results and Discussion

### 9.2.1 Structure and Phylogenetic Analysis of the Photosynthesis (PS) Operon.

The genome annotation of KT71 revealed the presence of a full PS superoperon (180) (KT71.19323–19518) on the smaller of the two large scaffolds [supporting information (SI) Fig. 4]. Both a smooth tetranucleotide signature (SI Fig. 5) (181) and the absence of transposons in the vicinity make it highly unlikely that the PS operon has been obtained by a recent lateral gene transfer. The operon consists of the typical subclusters *crtEF-bchCXYZ-puf* and *bchFNBHLM* (Fig. 9.2) but differs in the global and local arrangement from cultured alpha- and betaproteobacterial anoxygenic phototrophs (182). In these, the *puf* gene cluster coding for the light-harvesting complex I (LHC I) and the photosynthetic reaction center are usually arranged in the order of *pufBA-LMC* (183). In KT71, it is switched to *pufLMC-BA*. Interestingly, this gene arrangement is identical to that in two BAC clones, EBAC65D09 (AE008919) and EBAC29C02 (AE008920), retrieved from coastal bacterioplankton sampled at Moss Landing, CA (Fig. 9.2) (182). Further analysis also revealed identical arrangement for the *crt* and *bch* genes on both the BAC clones and KT71. A third BAC clone (EBAC69B03, GenBank accession no. AY458648) shares an identical arrangement of the *bchP-pucC-bchG-crtJ-bchFNBHLM-pucC-puhH* region with KT71. Based on comparative sequence analyses, these BAC clones were postulated to originate from *Gammaproteobacteria* (182). A gene-by-gene comparison of the BAC clones EBAC65D09, EBAC29C02, and EBAC69B03 with KT71 showed a high average sequence identity on the amino acid level of 56%, 55%, and 62%, respectively (Fig. 9.2; SI Table 1). We conclude that the three BAC clones 65D09, 69B03, and 29C02 indeed originate from *Gammaproteobacteria*, more precisely from members of the NOR5/OM60 clade.

Genes coding for a LHC II and a *pucC*-like transcriptional regulator were found clustered together on scaffold 1 (*pucBAC*: KT71.03072, KT71.03077, and KT71.03082). The LHC II complex proteins were most closely related to those of *Rhodospseudomonas palustris* (65% and 69% sequence identity in amino acids; best blast hit).



**Figure 9.3:** Pigment analysis. (A) Absorption spectrum of *Bchl a* extracted from *KT71*. Retention time, 15.16 min. (B) Absorption spectrum of spirilloxanthin-like carotenoid from *KT71*. Retention time, 17.33 min. Note: This curve was fitted (thick line) to better visualize the characteristic peaks.

### 9.2.2 Pigment Analysis

Cell extracts of *KT71* were subjected to spectrochromatographic analysis. Pigments found after consecutive acetone and methanol extractions followed by HPLC analysis showed the typical profiles for bacteriochlorophyll *a* (*Bchl a*) with the main peaks at 360, 580, and 776 nm (Fig. 9.3 A) and a carotenoid-like compound with absorption maxima at 470, 494, and 532 nm, respectively (Fig. 9.3 B). The latter absorption maxima are almost identical to those described for spirilloxanthin detected in *Roseateles depolymerans* (184). Significant amounts of *Bchl a* were detected only in cultures of *KT71* growing with light on the oligotrophic MPM-m (124) medium for an extended time (bacteriochlorophyll concentration of  $680 \frac{\mu\text{g}}{\text{liter}}$  after 4 mo). In contrast, *Bchl a* was never found in cultures grown to stationary phase in the nutrient-rich SYPG medium or in MPM-m medium without illumination. Genomic analysis showed that indeed all genes for the synthesis of spirilloxanthin are present. They are clustered together in the PS operon (*crtJ* and cluster *crtFECBI*), except for the *crtD* gene coding for a methoxyneurosporene dehydrogenase, which is found 300 kb separated on the first scaffold (*KT71\_07854*).

### 9.2.3 Photoautotrophy vs. Photoheterotrophy.

*KT71* has all the components necessary for a fully functional photosystem typical of anoxygenic phototrophs: the LHC I and II, a reaction center, and carotenoid pigments. Physiological tests indicated that *KT71* is not able to grow autotrophically. None of the key genes for autotrophic carbon fixation, like ribulose-1,5-bisphosphate-carboxylase/oxygenase (Calvin-cycle), ATP-citrate-lyase (reductive citrate cycle), or CO-dehydrogenase/acetyl-CoA-synthase (reductive acetyl-CoA pathway), were found in the genome, which is typical for aerobic anoxygenic phototrophs (AAnPs) (185). Most likely, *KT71* is able to gain energy from light by a light-dependent cyclic electron transport through the photosystem and the generation of a proton gradient (185). A proton-driven ATP synthase complex was annotated (*KT71\_04845–04885*). Alternatively, the proton gradient might be converted into a sodium gradient by proton/sodium antiporters, five of which have been found in the *KT71* genome (e.g., *KT71\_06212* and *KT71\_09322*). The sodium gradient in turn may drive a sodium-dependent ATPase (*KT71\_09367*) or may be directly used by the flagella motor (*KT71\_00645*).

First-growth experiments with *KT71* suggest an enhanced cell yield with light. Two flasks containing 960 ml of minimal MPM-m medium were inoculated with 4 ml of a stationary-phase culture and incubated for 4 wk at room temperature. From the culture grown with light from a 60-W light bulb, 32.4 mg of cell mass (dry weight) could be harvested, whereas from the parallel culture grown in the dark, only 17.6 mg of cell mass (dry weight) could be obtained. These experiments have to be regarded as prelimi-

nary, because no parallel experiments were done. Future experiments should also address starvation survival, because Breznak et al. (186) could show that the survival half-time of the facultative anaerobic anoxygenic phototroph *Rhodospirillum rubrum* (*Alphaproteobacteria*) was  $\approx 29$ -fold longer if grown with ambient-light intensities than without light.

#### 9.2.4 Putative Regulation of PS.

Annotation identified two genes containing a member of the sensor family BLUF (blue light using flavin adenine dinucleotide). One of the BLUF sensors was detected directly upstream of the PS superoperon. It contains the BLUF domain at the N terminus of the ORF KT71\_19323. In *Rhodobacter sphaeroides* BLUF forms part of the AppA protein, which regulates the expression of the PS cluster by sensing and integrating both the light and redox regimes (187). BAC clone EBAC29C02 also contains a BLUF sensor with a similar structure directly upstream of the PS operon suggesting an involvement of the BLUF sensor in the light regulation of the PS operon. Interestingly, in direct vicinity to the second BLUF sensor (KT71\_09447), a two component response regulator (KT71\_09452) could be found, suggesting, that this BLUF sensor forms part of a two component system (SI Table 2).

#### 9.2.5 Microaerophily.

KT71 is a strictly aerobic organism with a clear preference for low-oxygen niches. Typical enzymes necessary for the detoxification of oxygen, a bifunctional catalase-peroxidase (KT71\_02962) and a superoxide dismutase (KT71\_19732), could be annotated in the genome. It did not grow with nitrate as sole electron acceptor, nor was it able to ferment. No gene encoding a dissimilatory nitrate reductase was found in the genome. A putative sulfite/nitrite reductase-like enzyme (KT71\_15541) was annotated but is most likely involved in the assimilatory nitrate or sulfate reduction.

In deep-agar cultures, KT71 forms distinct bands several millimeters below the surface. The position of the visible cell layer depended on the substrate concentration in the medium and was closer to the surface at higher substrate concentrations. To determine the exact oxygen concentration for optimal growth of KT71, oxygen profiles were measured in cultures grown in SYPG medium with 0.15% (wt/vol) agar (soft agar). An oxygen profile measured with microsensors from the surface of the soft agar down to a depth of 8 mm is shown in SI Fig. 6. The highest cell density was visible at an oxygen saturation of  $\approx 10\%$  ( $30 \mu\text{M O}_2$ ). Experiments with varying substrate concentrations showed that KT71 exhibits an excellent chemotaxis for suboxic oxygen conditions. KT71 is motile and possesses a complete flagellum operon (gene loci KT71\_00565–00780). Next to the *aa3*-type terminal cytochrome *c* oxidase (KT71\_04625–04640), KT71 harbors a *cbb3*-type cytochrome *c* oxidase (*fixNOQP*, KT71\_16991–17006). Such terminal cytochrome *c* oxidases with high oxygen affinity are expressed only under reduced oxygen conditions in *Bradyrhizobium japonicum* (188). In *R. sphaeroides*, the same enzyme complex was shown to be involved in the signal transduction and functions as a redox sensor (189) (SI Table 2), which might be also the case in KT71.

#### 9.2.6 Substrate Spectrum.

Substrate utilization tests indicate that KT71 prefers complex substrate mixture for growth (e.g., yeast extract or Trypticase peptone), whereas many monomeric substrates given as sole source of carbon and energy are used a little or not at all. As an exception, KT71 can grow well with carbon sources such as glutamate, pyruvate, and fatty acids, most likely due to the fact they play central roles in the metabolism of this organism. Glutamate is a central metabolite and is presumably taken up by a proton/sodium-glutamate symport protein (KT71\_01885). It is further fed by two glutamate dehydrogenases into

the trichloroacetic acid cycle (KT71\_16246 and KT71\_18661) or into the proline synthesis pathway (glutamate-5-kinase, KT71\_02697). Pyruvate is presumably being metabolized by a pyruvate-dehydrogenase (KT71\_00115) and further metabolized by the citric acid cycle.

Annotation identified all genes necessary to perform the complete pentose phosphate pathway. This pathway plays a central role in the anabolism of nucleotides and amino acids as well as the generation of reducing power by NADPH synthesis. Laboratory experiments have shown that KT71 is not able to use glucose as sole source of carbon and energy. Alonso and Pernthaler (190) could not detect any glucose uptake of NOR5/OM60 *in situ* under both oxic and anoxic conditions in the North Sea for the entire NOR5/OM60 clade. In the genome, all genes for glycolysis are present, except for the initial activating enzymes. Neither a glucose phosphorylating glucokinase nor an intact phosphotransferase system (PTS) was found. For the latter, only the specific phosphocarrier HPr (KT71\_10197) and a single-chain EIIA of the PTS could be annotated (KT71\_10207).

The genome contains several genes coding for putative lipase/esterases and proteases/peptidases that might be involved in the breakdown of lipids and peptides. In the laboratory, no hydrolysis of the polysaccharides starch, cellulose, or chitin by KT71 could be detected, in line with the annotation of the genome. A lipase/esterase activity could be confirmed by the hydrolysis of the artificial substrates Tween 80 (Polyoxyethylenesorbitan monooleate) and Tween 20 (Polyoxyethylenesorbitan monolaurate). Gelatin and casein were tested negatively as potential substrates for proteases and peptidases. Although proteases can have a high specificity for distinct substrates, this finding points to a preferred utilization of oligopeptides or partly degraded proteins by KT71. Two transporters for oligopeptides with up to five amino acids were found, *oppABC* (KT71\_06839–06854) and *oppF* (KT71\_00435). Culture experiments show that KT71 is able to synthesize all essential amino acids and most of the vitamins, except for biotin, thiamin, and vitamin B12. Two TonB-dependent vitamin B12 sensors (KT71\_17391 and KT71\_18621) and an ABC vitamin B12 transporter system *btuCDF* (KT71\_17396–17411) were found in the genome. The annotation is consistent with this auxotrophy and the inability to use many substrates (e.g., glucose) as single sources of carbon and energy.

### 9.2.7 Storage Compounds.

Two highly similar genes coding for cyanophycin synthetases were found in tandem (KT71\_18591 and KT71\_18596; 38% identical amino acids). Cyanophycin synthetase is described as a homodimer but was also considered to form heterodimers of the type CphA and CphA' (191). Both genes have high similarity to the cphA genes in the Gammaproteobacteria *Colwellia psychoerythraea* 34H and *Francisella tularensis* (59% and 56% identity for the long CphA and 33% and 30% amino acid identity for the short CphA', respectively). Cyanophycin is a polymer of aspartic acid and arginine. It was first found as a storage compound in cyanobacteria and subsequently detected in many heterotrophic bacteria. The polymer forms insoluble granula inside the cell that can be extracted with diluted acids (192). Cells containing highly refractile granulas could be mainly observed in stationary cultures of KT71 grown under conditions of a high ratio of nitrogen to carbon. Cyanophycin was identified in these granula by a negative reaction with the lipophilic stain, Nile blue A, and dissolution in diluted HCl (see SI Fig. 7). A cyanophycinase was not annotated, but most likely the polypeptide is degraded by an unknown peptidase. The formation of polyphosphate is not yet confirmed by physiological tests but two enzymes, an inorganic polyphosphate/ATP-NAD kinase (KT71\_14354) and a polyphosphate kinase (KT71\_16696), were found in the genome.

### 9.2.8 Formation of Aggregates and Polysaccharide Production.

In pure cultures of KT71, the formation of large flocs was observed (SI Fig. 8A). There is microscopic evidence that members of the NOR5/OM60 clade attaches also in nature to macroscopic particles (SI Fig. 8B). Genome analyses revealed several features consistent with aggregation. Several loci in the KT71 genome code for the synthesis of type IV pili or fimbriae (193). The formation of pili seems to be regulated by a sensory mechanism encoded by the genes *pilS* (KT71\_19657) and *pilR* (KT71\_19662; SI Table 2). In addition, two operons were found containing exopolymer producing proteins (KT71\_09752–09807 and KT71\_06404–06469). These operons comprised genes for polysaccharide length-determinant proteins [KT71\_09767 and KT71\_06439], (exo)-polysaccharide biosynthesis protein (KT71\_09772 and KT\_066454), polysaccharide polymerases (KT71\_009807), polysaccharide export proteins (KT71\_09762 and KT71\_06444)], and some glycosyltransferases (e.g., KT71\_06459, KT71\_06434, KT71\_09787, and KT71\_06429). Interestingly, each of the operons contains a two-component sensor kinases/response regulator (SI Table 2). Based on the current annotation, it is not clear to which stimuli they respond.

### 9.2.9 Sulfur Metabolism.

KT71 most likely uses the APS/PAPS pathway to obtain reduced sulfur compounds. Three genes coding for the key enzymes of that pathway were annotated in KT71, a sulfate adenylyltransferase (KT71\_10572), an adenylylsulfate kinase (KT71\_10567), and a phosphoadenosine phosphosulfate reductase (KT71\_06329). Genome annotation revealed that the gene cluster *soxH-RCDXYZA-B* is potentially involved in the oxidation of reduced sulfur compounds (KT71\_03447–03482 and KT71\_03497). This cluster contains the core gene set *soxXYZAB*, which is found in many species capable of oxidizing reduced sulfur compounds (194). A comparison with other sulfur-oxidizing organisms shows that the gene arrangement *soxH-RCDXYZA-B* is unique to KT71 and has not been found in any of the species described to date. Unlike in *Silicibacter pomeroyi* (130), the supplementation of media with the inorganic sulfur compounds thiosulfate or sulfur did not significantly promote growth of KT71 in cultivation experiments using different carbon sources. The inability to gain additional energy by the oxidation of reduced inorganic sulfur compounds may be due to the lack of several *sox* genes compared with the exemplary cluster of *sox* genes found in the genome of *Paracoccus pantotrophus* or *S. pomeroyi* (130). Of special interest is the lack of the gene *soxV* that codes in *P. pantotrophus* GB17 for a membrane protein that is predicted to transfer electrons from the cytoplasm to the periplasmic thioredoxin *soxW* (195). It was shown that inactivation of SoxV in *P. pantotrophus* and the phototrophic bacterium *Rhodovulum sulfidophilum* leads to a phenotype that is unable to use thiosulfate for energy conservation (196; 197). Despite this finding, the possibility exists that KT71 can use alternative sulfur compounds like dimethylsulfoniopropionate or dimethylsulfide that were not tested yet. These compounds are present in high amounts after algal blooms and are metabolized by *Roseobacter* species (198; 199).

### 9.2.10 *C. litoralis*: A Typical Shelf Bacterium?

KT71 was isolated from coastal surface water in the rather shallow German Bight. There, the water column is close to oxygen saturation during most of the year. It came as a surprise that KT71 avoids sites with oxygen saturation and grows optimally under microaerobic conditions. In coastal areas, suboxic conditions are found, temporarily, in large macroscopic aggregates (200), and permanently a few millimeters below the sediment surface (201). Therefore, we hypothesize that the habitat range of KT71 includes particles and sediment surfaces.

Shallow shelf areas are characterized by extensive mixing of the water, sediment interface by tides, or wind stress. Resuspension of sediment particles into the water

column is followed by periods of sedimentation. Thereby, in temperate coastal systems like the German Bight, marine microorganisms are faced with pronounced fluctuations of multiple parameters such as substrate, nutrient, and oxygen concentrations, as well as light levels on a daily and seasonal scale. Based on our genomic and ecophysiological data, KT71 seems well adapted to such a dynamic shallow shelf environment.

Organic particles are nutrient-rich hotspots in the otherwise oligotrophic water column (202). By attaching to their surfaces, KT71 may directly use mono- and oligomeric substrates or may benefit, as a commensal, from the exoenzymatic activities of polymer-degrading bacteria such as *Rhodopirellula baltica* (21) and *Gramella forsetii* (61). The possibility that KT71 is actively shaping its environment by facilitating the formation of “marine snow” by polysaccharide production needs to be addressed in future studies. Particle association also serves as a transport mechanism to the sediment surface. There, KT71 may thrive on low-molecular-weight substrates like peptides or lipids that accumulate on the sea floor (203).

The presence of a complete mercury-resistance operon (KT71\_16196–16226) in the genome of KT71 is consistent with a prevalence of this strain in the suboxic zone of sediments. It is known that low-redox potentials and the degradation of complex organic matter in upper sediment layers lead to the mobilization of active mercury species in the form of inorganic ions ( $\text{Hg}^{2+}$ ) or weak inorganic complexes (see, e.g., ref. (204)). In a recent study, depth profiles of reactive mercury species were determined in North Sea sediment, and it was found that peak values are reached at the sediment water interface (205). Hence, genes that confer resistance to toxic mercury ions may be much more important for bacteria dwelling in the surface sediment than for bacteria indigenous to the water column.

A specific highlight of KT71 is the presence of a PS superoperon. It is becoming more clear that photoheterotrophy is widespread among marine microorganisms (182; 206). By the light-driven generation of a proton gradient, KT71 might be able to survive extended periods of starvation, e.g., during the winter period. The storage compounds cyanophycin and polyphosphate are yet another adaptation to famine situations. Interestingly, in contrast to all other AAnPs known to date, KT71 produces only trace amounts of carotenoids and shows no obvious pigmentation (SI Fig. 8c). Because a major function of carotenoids is the protection of cells from damage by UV radiation, this may reflect an adaptation to low-light zones, i.e., depths of several meters in the water column or subsurface sediment layers in shallow water. Recently, strongly pigmented strains closely related to KT71 were isolated from surface sediments in the Wadden Sea, suggesting the ability of members of the NOR5/OM60 clade to adapt also to high light conditions (J. Harder, personal communication). These strains will also allow us to determine whether PS is a general feature of the NOR5/OM60 clade.

### 9.2.11 Significance of the NOR5/OM60 Clade Represented by KT71.

It has been estimated that AAnPs account for  $\geq 10\%$  of the bacterioplankton community in the oligotrophic open ocean (207; 208; 209). More recent studies have shown that AAnPs may be less important in the open ocean ( $\approx 1\%$ ) (210) but can reach up to 15% abundance in eutrophic and mesotrophic coastal areas (211; 212). Currently, the alphaproteobacterial *Roseobacter* clade is considered to be the dominant group of marine AAnPs (199; 213; 214). The discovery of BAC clones with PS operons showing best BLAST hits to *Gammaproteobacteria* (182) clearly suggested the existence of a second group of AAnPs. The genome analysis of *C. litoralis* strain KT71 identifies this microorganism as a cultured representative of the gammaproteobacterial marine AAnPs, enabling future experiments investigating the importance of gammaproteobacterial AAnPs in coastal environments by using KT71 as a model organism.



## 9.3 Materials and Methods

### 9.3.1 Sequencing and Assembly.

Sequencing of KT71 was done in a conventional whole-genome shotgun sequencing approach. Two genomic libraries with insert sizes of 4 and 40 kb were made as described in Goldberg et al. (215). The prepared plasmid and fosmid clones were sequenced from both ends to provide paired-end reads at the J. Craig Venter Science Foundation Joint Technology Center on Applied Biosystems 3730XL DNA sequencers (Applied Biosystems, Foster City, CA). Whole-genome random shotgun sequencing produced 38,544 good reads averaging 892 bp in length, for a total of  $\approx 34.38$  Mbp of microbial DNA sequence ( $=7.53 \times$  coverage).

The genome of KT71 contains  $\approx 4.36$  Mb with an average GC content of 57.7% and is deposited under GenBank accession no. AAOA00000000. Successful reads for each organism were used as input for the Celera Assembler (216). Data are released to the GBMF Marine Microbial Genome Sequencing Project web site (<https://research.venterininstitute.org/moore>) and GenBank. A genome report compliant with the “Minimum Information about a Genome Sequence specification” is available from the Genome Catalog at <http://www.genomics.ceh.ac.uk/genomecatalogue>.

### 9.3.2 Analysis of the Genome Architecture.

The cumulative GC-skew ( $[(G-C)/(G+C)]$ ) was computed with a custom PERL script (SI Fig. 9). Genome-wide fluctuations of the GC-skew, AT-skew, GC-content, DNA curvature and DNA bending, (SI Fig. 4) were computed with custom PERL scripts and the programs banana and btwisted from the EMBOSS suite (217), respectively, and visualized with GeneWiz (218). The positions of all genes and subsets of related genes (SI Fig. 4) were visualized with the program GenomePlot (219). Genome-wide fluctuations in tetranucleotide composition (SI Fig. 5) were calculated and plotted with the program TETRA (181).

### 9.3.3 Gene Prediction and Annotation of the Genome Sequence.

Potential protein-coding genes were identified by the Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP) of the National Center for Biotechnology Information. PGAAP combines three distinct gene finders, GenMark (220), GenMark.hmm (221), and Glimmer2 (222), and parses the individual results using a conflict-resolving strategy. BLAST analysis of the intergenic regions resulted in finding one additional short ORF (*pujB*, KT71\_19424). Transfer RNA genes were identified by using tRNAScan-SE (163), and ribosomal RNA genes were detected by standard BLAST similarity searches (7) against public nucleotide databases.

Annotation of the genome sequence was performed with the GenDB v2.2 annotation system (9). For each ORF, similarity searches against various sequence databases (NCBI nr, NCBI nt, and SwissProt) and protein family databases (Pfam, Prosite, InterPro, and COG) were performed. In addition, potential signal peptides were predicted with SignalP Ver. 2.0 (165), and potential transmembrane helices were predicted with TMHMM Ver. 2.0 (63). Based on this information, all ORFs were automatically annotated in a fuzzy logic-based approach (10). This automatic annotation was extensively manually checked and refined for each ORF. Predicted protein-coding genes were functionally classified according to COG Ver. 3 (223) (see SI Text).

### 9.3.4 Comparative Sequence Analyses.

16S rRNA gene sequences related to KT71 were downloaded from the National Center for Biotechnology Information and imported into a 16S rRNA database. Phylogenetic

reconstructions were done with the ARB package ([www.arb-home.de](http://www.arb-home.de)) (1) by using maximum parsimony, maximum likelihood, and neighbor-joining methods with different filters and matrices (see SI Text).

### 9.3.5 Physiological Tests.

KT71 was routinely grown and maintained either in the oligotrophic MPM-m medium described by ref. (124) (see SI Tables 3–7 for details) or in the complex medium SYPG containing the following compounds per liter of distilled water: 35.0 g of sea salts, 0.5 g of yeast extract, 0.25 g of Trypticase peptone, and 0.1 g of sodium l-glutamate. Utilization of substrates was tested in a mineral medium containing (per liter) 35.0 g of sea salts, 0.1 g of NH<sub>4</sub>Cl, 0.05 g of K<sub>2</sub>HPO<sub>4</sub>, and 10 ml of a vitamin solution (see DSMZ medium 141, [www.dsmz.de](http://www.dsmz.de)). Standard tests for the detection of enzymes like catalase, oxidase, lipase/esterase, and proteases were done according to the protocols given in ref. (224). Oxygen measurements in soft-agar medium were done with a Clark-type oxygen microelectrode, as described (225).

### 9.3.6 HPLC Analysis of Photosynthetic Pigments.

Pigments from cell pellets of KT71 were obtained by freeze drying and consecutive acetone and methanol extraction with sonication. HPLC analyses of cell extracts were done on a Waters 2690 Separation Modul (Waters, Eschborn, Germany) with a 250 × 4.6-mm vortex column packed with Eurospher-100 C 18 (particle size, 5 μm; Knauer, Berlin, Germany) and a Waters 996 Photodiode Array Detector. The mobile phase was chosen after Wright and Jeffrey (226). Bchl<sub>a</sub> was identified by retention-time coinjection of a Bchl<sub>a</sub> standard from Rhodospseudomonas sphaeroides (Sigma–Aldrich, Taufkirchen, Germany) and spectrography at 384 nm. The quantity of Bchl<sub>a</sub> was estimated from the areas under the peaks, which were calibrated with the Bchl<sub>a</sub> standard. No standard was available for the spirilloxanthin-like compound, but the carotenoid was identified by comparison with similar spectra at similar retention times from, e.g., ref. (227).

## 9.4 Acknowledgments

Jakob Pernthaler initially suggested sequencing KT71 in the framework of the Marine Microbiology Sequencing Initiative of the GBMF. Sequencing, assembly, and annotation efforts were supported by the GBMF as part of its Marine Microbial Sequencing Project ([www.moore.org/marinemicro](http://www.moore.org/marinemicro)). We thank Granger Sutton and his team for the ongoing development and maintenance of the Celera Assembler and related tools and Aaron Sutton for help with specific analysis issues. The JCVI software team, under the leadership of Saul A. Kravitz, manages assembly, annotation, and Web delivery of data for the GBMF-funded project (<http://research.venterininstitute.org/moore>). We thank Robert Friedman for his leadership of the GBMF-funded project. Many thanks for help with pigment extraction and HPLC analysis to Gabriele Klockgether and Raphaela Shoon. Armin Gieseke is acknowledged for help with the microsensor measurements. Jens Harder shared unpublished results on strongly pigmented strains closely related to KT71. We are indebted to Gunnar Gerdts and Antje Wichels from the Alfred Wegener Institute for Polar and Marine Research for constant support for field work on Helgoland. We thank Oded Beja, Margarete Schüler, Marcel Kuypers, and Anke Meyer-dierks for helpful suggestions and discussions. We acknowledge the J. Craig Venter Institute (JCVI) Joint Technology Center, under the leadership of Yu-Hui Rogers, for producing the genomic libraries and the sequence data. This project was funded by

the GBMF, the Max Planck Society, and the FP6-EU Network of Excellence Marine Genomics Europe (Grant GOCE-CT-2004-505403).



## Chapter 10

# Pirellula Paper

### Transcriptional response of the model planctomycete *Rhodopirellula baltica* SH1<sup>T</sup> to changing environmental conditions.

P. Wecker<sup>a,c</sup>, C. Klockow<sup>a,c</sup>, A. Ellrott<sup>a</sup>,  
C. Quast<sup>a</sup>, P. Langhammer<sup>b</sup>, J. Harder<sup>b</sup>,  
F.O. Glöckner<sup>a,c</sup>

<sup>a</sup>Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>b</sup>Department of Microbiology, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>c</sup>Jacobs University Bremen gGmbH, D-28759 Bremen, Germany;

**Journal; Volume (Issue):** BMC Genomics; 10 (1)  
**Pages:** 410  
**Month / Year:** September 2009  
**DOI:** 10.1186/1471-2164-10-410

#### Contributions:

Automatic re-annotation of the genome of *Rhodopirellula baltica* SH1<sup>T</sup> and general Bioinformatics support. Concept, design and implementation of the AJAX / GWT based genome browser provided for inspecting the genome (<http://gendb.mpi-bremen.de/gendb/BX119912>).

---

**Contents**

---

<b>10.1 Background</b>	<b>121</b>
<b>10.2 Results and Discussion</b>	<b>122</b>
10.2.1 Overview	122
10.2.2 Experimental design and array data quality assessment	122
10.2.3 Effect of stress on <i>Rhodopirellula baltica</i>	123
10.2.4 Specific results of the shift experiments	123
10.2.5 Common stress response	127
10.2.6 Hypothetical proteins	128
10.2.7 <i>Planctomycete</i> special feature: Genes encoding sulfatases	129
<b>10.3 Conclusion</b>	<b>129</b>
<b>10.4 Methods</b>	<b>130</b>
10.4.1 Bacterial growth conditions	130
10.4.2 Sample collection, cell lyses, RNA Isolation and cDNA synthesis	130
10.4.3 Experimental design and sample preparation	130
10.4.4 Heat shock from 28°C to 37°C	131
10.4.5 Cold shock from 28°C to 6°C	131
10.4.6 Salt stress from 17.5‰ to 59.5‰ salinity	131
10.4.7 Whole Genome Array construction, hybridization and image analysis	131
10.4.8 Signal detection and data analysis	131
10.4.9 Cluster analysis	132
10.4.10 Genome tools	132
10.4.11 Microarray Datasets	133
<b>10.5 List of abbreviations</b>	<b>133</b>
<b>10.6 Competing interests</b>	<b>133</b>
<b>10.7 Authors contribution</b>	<b>133</b>
<b>10.8 Acknowledgements</b>	<b>133</b>

---

## Abstract

**Background:** The marine model organism *Rhodopirellula baltica* SH1<sup>T</sup> was the first *Planctomycete* to have its genome completely sequenced. The genome analysis predicted a complex lifestyle and a variety of genetic opportunities to adapt to the marine environment. Its adaptation to environmental stressors was studied by transcriptional profiling using a whole genome microarray.

**Results:** Stress responses to salinity and temperature shifts were monitored in time series experiments. Chemostat cultures grown in mineral medium at 28°C were compared to cultures that were shifted to either elevated (37°C) or reduced (6°C) temperatures as well as high salinity (59.5‰) and observed over 300 min. Heat shock showed the induction of several known chaperone genes. Cold shock altered the expression of genes in lipid metabolism and stress proteins. High salinity resulted in the modulation of genes coding for compatible solutes, ion transporters and morphology. In summary, over 3000 of the 7325 genes were affected by temperature and / or salinity changes.

**Conclusions:** Transcriptional profiling confirmed that *R. baltica* is highly responsive to its environment. The distinct responses identified here have provided new insights into the complex adaptation machinery of this environmentally relevant marine bacterium. Our transcriptome study and previous proteome data suggest a set of genes of unknown functions that are most probably involved in the global stress response. This work lays the foundation for further bioinformatic and genetic studies which will lead to a comprehensive understanding of the biology of a marine *Planctomycete*.

## 10.1 Background

Marine ecosystems, covering approximately 71% of the Earth's surface, host the majority of biomass and contribute significantly to global cycles of matter and energy. Microorganisms are known to be the 'gatekeepers' of these processes, and insight into their lifestyle and fitness enhances our ability to monitor, model and predict the course and effect of global changes. Nevertheless, specific knowledge about their functions is still sparse. The 'genomic revolution' (228) has opened the door to investigations targeting their genetic potential and activity on the molecular level.

A particularly interesting representative of the marine picoplankton community is *Rhodopirellula baltica* SH1<sup>T</sup>, a free-living bacterium which was isolated from the water column of the Kiel Fjord (Baltic Sea) (229). *R. baltica* belongs to the phylum *Planctomycetes*, a broadly distributed group of bacteria, whose members can be found in terrestrial, marine and freshwater habitats (115; 230; 231; 232; 169), but also in extreme environments like hot springs (233), marine sponges (234) and the hepatopancreas of crustaceans (235).

In terms of cell biology all *Planctomycetes* share several morphologically unique properties, such as a peptidoglycan-lacking proteinaceous cell wall (236; 237), intracellular compartmentalization (238) and a mode of reproduction via budding. The latter results in a cell cycle that is characterized by motile and sessile morphotypes similar to *Caulobacter crescentus* (239; 240; 21; 241). A specific holdfast substance produced by sessile cells allows *R. baltica* to attach to macroscopic detrital aggregates (marine snow) (115; 169).

At present, four planctomycete genomes are currently available (75). Of these, the genome of *R. baltica* is the only one completely closed (21). The genome was found to be 7,145,576 bases in size and codes for 7325 open reading frames (ORFs) plus 72 RNA genes. Originally, only 45% of the ORFs were assigned particular functions (21). Thus, over 55% of all proteins in the genome remain functionally uncharacterized. These were referred to as 'hypothetical proteins' with or without the affix 'conserved' contingent on wider phylogenetic distribution (242). A subset of these conserved hypothetical proteins is specific for *Planctomycetes* (75). It seems likely that some of these genes code for the unique planctomycetal cellular characteristics and metabolic traits.

The availability of the genome information triggered several key post-genomic studies including studies of the proteome (243; 244; 245; 246; 247), enzyme activity (248) and protein crystallization (249).

In summary, these studies confirmed the hypothesis of Glöckner et al. that *R. baltica* is a polysaccharide degrader (21). It appears *R. baltica* is gaining carbon and energy from the decomposition of complex heteropolysaccharides originally produced by algae in the photic zone while slowly sedimenting with the marine snow.

Marine microorganisms like *R. baltica* are exposed to rapidly changing environmental conditions such as varying temperature, salinity, irradiance and oxygen concentration. Typically, sudden changes of these environmental conditions induce a stress response in the exposed planktonic community characterized by a distinct change in their gene expression pattern. This stress response enables the organisms to protect vital processes and to adapt to the new condition. Such responses have been described for a set of organisms from different environments including *Shewanella oneidensis* (250; 251), *Pseudomonas aeruginosa* (252), *Desulfovibrio vulgaris* Hildenborough (253), *Xylella fastidiosa* (254), *Synechocystis sp.* (255), and *Yeast* (256).

To gain insights into the stress responses of *R. baltica* with respect to salinity and temperature the first whole genome array for *R. baltica* - also the first planctomycete microarray - was established and applied. The reported data will serve as a resource to expand our understanding of the physiological and transcriptional response of *R. baltica* to the wide range of changing environmental conditions a free-living marine bacterium is exposed to.

## 10.2 Results and Discussion

### 10.2.1 Overview

54 distinct, total RNA samples were analyzed by whole-genome microarray hybridization. Differential expression of 2372, 922 and 1127 genes was noted during heat shock, cold shock and salt stress respectively at one or more of the five time points when compared to reference samples (Figure 10.1 i; ii & iii). With only 45% of the genes in *R. baltica*'s genome functionally annotated, it is not surprising that most of the differentially expressed genes were hypothetical or conserved hypothetical proteins. The complete list of the differentially expressed genes for each shift experiment and time point is available in the ADDITIONAL FILE 1.

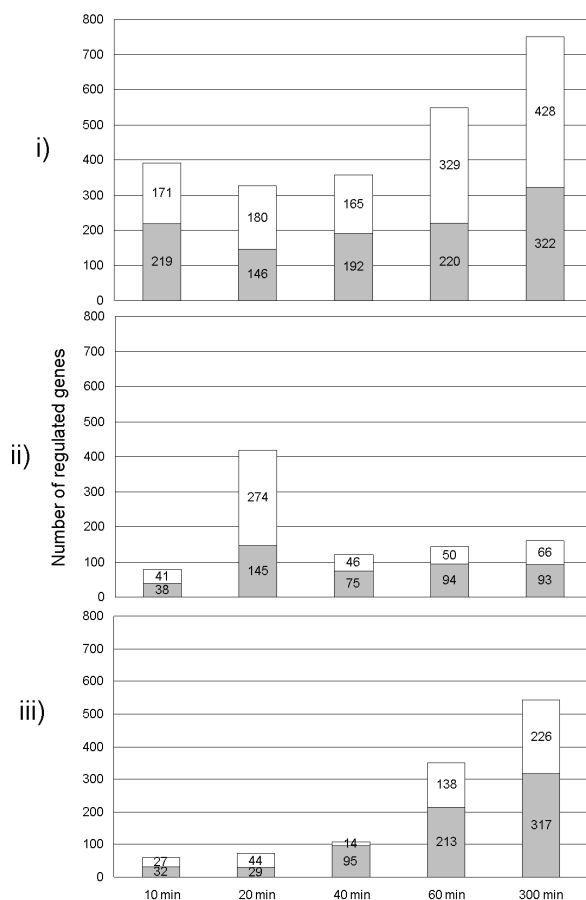
Only 32% of the regulated genes in the heat and cold shock experiments could be assigned with a COG function (Figure 10.2 i & ii) while 37% were assignable in the salt stress experiment (Figure 10.1 iii). This is in line with the 36% (2661 genes) of COG functional class designations in *R. baltica*. A striking feature of the expression profiles displayed is the stereotypical response of a large fraction of the genome to all three stress conditions. In summary, 152 genes are up- or down-regulated at any time point for all stressors. Of these 152 genes, 62 are induced and 90 are repressed (Table 10.1 and Table 10.2). 49% of the induced and 61% of the repressed genes were annotated as hypothetical proteins. The Venn diagrams shown in Figure 10.3 provide an overview of the specific and common genes of the three stress-specific responses. To identify co-regulated patterns of gene expression, we classified all differentially expressed genes of all three stress expressions into 30 k-means clusters based on their expression log ratio. To determine the necessary number of clusters a figure of merit was generated. 30 clusters were considered as adequate. The cluster data are available in the ADDITIONAL FILE 2. Clusters 1, 3 and 4 show a similar response to the specific environmental changes, called environmental stress response (ESR) over all experiments. Clusters 2, 4, 5, 7, 15 and 22 describe genes reacting to a specific environmental factor.

### 10.2.2 Experimental design and array data quality assessment

The experimental conditions used were chosen to mimic the natural environment of *R. baltica*; however, stress conditions were constrained by the detection limit of the microarray technology used and, hence, were required to elicit a sufficiently pronounced response from the organism. In contrast to steady-state or single-time-point studies, time series experiments can show the dynamic of gene expression.

The negative, positive and stringency controls printed on the array gave no indications for unspecific hybridizations. Co-hybridizations of two cDNA samples prepared from the same total cellular RNA (self-self hybridization) suggested that genes with an expression log ratio value greater than 1.5 and smaller than -1.0 for heat and cold shock, respectively, could be regarded as differentially expressed. Salt stress log ratio values over 1.2 and below -1.0 were considered as significant.





**Figure 10.1:** Number of regulated genes per stress experiment. Columns show the total number of up- (gray) and down- (white) regulated genes at each sampled time point compared to reference samples. i) heat shock, ii) cold shock and iii) high salinity

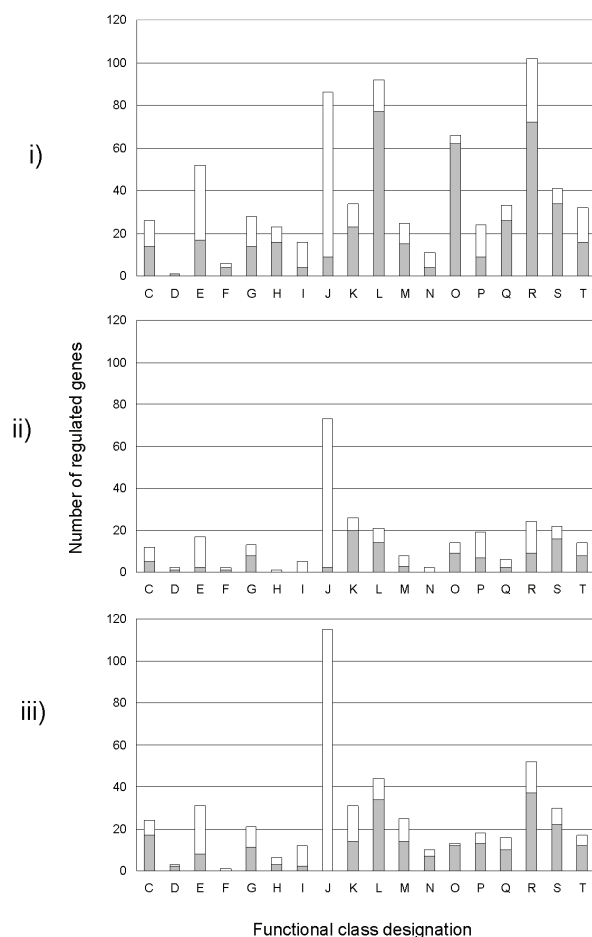
### 10.2.3 Effect of stress on *Rhodospirellula baltica*

No growth was detectable during stress conditions nor were any obvious morphological changes by microscopic investigation. Under optimal conditions *R. baltica* has a doubling time of 10-12 hours (244), suggesting physiological effects are not measurable during the short stress period of, at maximum 5, hours.

### 10.2.4 Specific results of the shift experiments

#### Heat shock

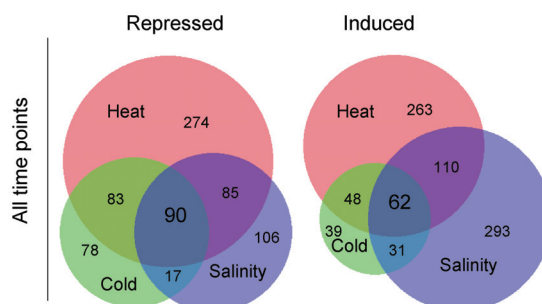
In their natural environment *R. baltica* cells can be regularly exposed to higher temperatures, for example, due to irradiation at the water surface. Therefore, *R. baltica* cells were rapidly shifted from 28°C to 37°C and observed over a period of 300 min in the first experiment. This is approximately 9°C above the optimal growth temperature reported by Schlesner et al. (229). Employing a higher temperature is very likely to kill the cells. The time series reveals a quick response of *R. baltica* to sudden temperature up-shifts. In total 2372 genes are regulated out of which 1140 genes encode hypothetical proteins. 390 genes (5%) were regulated after 10 min. This number increased to 750 genes (10%) after 300 min (Figure 10.1 i). The COG classes containing the translation [J] and amino acid transport and metabolism [E] were the largest down-regulated classes. Up-regulated genes were assigned to the COG classes of replication, recombination and repair [L], post-translation modification, protein turnover and chaperons



**Figure 10.2:** Number of regulated genes with an assigned COG-category. Columns show the number of up- (gray) and down- (white) regulated genes per assigned COG-category according to the NCBI database (cut off  $e$ -value  $e^{-4}$ ).

i) heat shock, ii) cold shock, and iii) high salinity;

Columns: [C] Energy production and conversion, [D] Cell division and chromosome partitioning, [E] Amino acid transport and metabolism, [F] Nucleotide transport and metabolism, [G] Carbohydrate transport and metabolism, [H] Coenzyme metabolism, [I] Lipid metabolism, [J] Translation, ribosomal structure and biogenesis, [K] Transcription, [L] DNA replication, recombination and repair, [M] Cell envelope biogenesis, outer membrane, [N] Cell motility and secretion, [O] Post-translational modification, protein turnover, chaperones, [P] Inorganic ion transport and metabolism, [Q] Secondary metabolites biosynthesis, transport and catabolism, [R] General function prediction only, [S] Function unknown, [T] Signal transduction mechanisms.



**Figure 10.3:** Venn diagrams of specific and common stress response. The diagram shows the distribution of stimulus-specific and common stress responses. All genes of all time points are represented in one diagram divided into repressed and induced genes.

[O], transcription [K], secondary metabolites biosynthesis, transport and catabolism [Q], cell envelope biogenesis, outer membrane [M] and general function prediction [R] (Figure 10.2 i).

Taking a closer look at the response of *R. baltica* to thermal stress revealed the induction of many known heat shock proteins (Hsp): ClpB (RB6751), GroEL (RB8970), DnaJ (RB8972), GrpE (RB8974), Hsp20 (RB10279, RB10283), *dnaK* (RB9105), as well as the ATP-dependent protease ClpP (RB9103). Also up-regulated were the chaperonins Cpn10 (RB10627 and RB8969) and Cpn60 (RB8966) as well as the cell division protein FtsH (RB2966) (Cluster 4 in ADDITIONAL FILES 2). Previous proteomic studies found the proteins of these genes as well, except FtsH, DnaJ and Hsp20 (244; 246).

The regulation of the heat shock response in *R. baltica* involves many transcriptional regulators. TetR (RB838) and GntR (RB1862, RB8695) showed an up-regulation, which affirms their important role in early heat shock response (257). A gene encoding for GntR was also found in the environment on the planctomycete fosmid 3FN from a Namibian coast metagenome study (75). In *E. coli* the induction of the majority of heat shock genes results from a rapid and transient increase in the cellular level of an alternative 32-kDa sigma factor (sigma32) encoded by *rpoH* along with the alternative sigma factors E and 54, encoded by *rpoE* (RB2302) and *rpoN* (RB6491), respectively (249). Although, all genes are present in the *R. baltica* genome, they were not observed to be regulated, suggesting a significantly different response cascade.

*R. baltica* also showed an extracytoplasmic stress response. The gene coding for SecA (RB11690), belonging to the Sec system, was induced. This indicated an activation of protein translocation, most probably from the riboplasma to the paryphoplasm or medium. Proton channels were induced and motility was inhibited as the flagellar motor switch protein (FliG - RB12502) was down-regulated after 20 min. This was followed by the inhibition of the type 4 fimbrial assembly protein (*pilC* - RB11597) after 40 min.

## Cold Shock

To investigate the response to cold shock, *R. baltica* cells were shifted from the optimal growth temperature 28°C (258) to 6°C and observed for a period of 300 min. 6°C was chosen for this study as this is a common temperature in the Baltic Sea. Sudden temperature changes occur naturally due to turbulences between water layers. Further, the temperature difference of 22°C is generally regarded as standard for cold shock studies with bacteria (250; 259). Compared to heat shock only one third (922) of the regulated genes were differentially expressed. Out of these 922 regulated proteins, 391 genes (42%) encode for hypothetical proteins. The cold shock response reached its peak after 20 min with 419 differentially expressed genes (6%) and decreased thereafter (Figure 10.1 ii). In contrast to the heat shock experiment, it seemed that *R. baltica* needed approximately one hour to adapt to cold conditions. Like other bacteria, *R. baltica* responded to cold conditions with the up-regulation of genes coding for stress response [COG class O], cell envelope and transport [M], transcription factors and solute uptake. Genes for amino acid biosynthesis [E] as well as protein fate and synthesis [J] were down-regulated (Figure 10.2 ii) (251).

Transcriptional activity was regulated by the up-regulation of diverse RNA polymerase

sigma factors, such as *rpoD* (RB6780) and *sigK* (RB1392). A homolog of *rpoD* (RB6780) was also found on the planctomycete fosmid 13FN (75). 20 min after the exposure of *R. baltica* to cold stress conditions it started to express genes implicated in the modification of cytoplasmic membrane composition, fluidity as well as morphology. The alteration of the lipid composition in the cold has been previously reported in other microorganisms (260). In *R. baltica* genes coding for cell envelope (RB6114 and RB6895), transport (RB4870), lipid metabolism (RB316) and 18 genes coding for membrane proteins were repressed after 20 min.

Furthermore, *R. baltica* repressed genes involved in sporulation *oppB* (RB12861) and O-antigen flippase (RB2503), *flaA* (RB4454) and pilus assembly (RB4061 and RB5478), leading to reduced motility and budding ability. Genes associated with amino acid biosynthesis, especially with synthesis and fate of glutamine (RB4269) and glutamate (RB5653) were also affected. The latter have been shown to be translated (245; 247). A glycosyltransferase (RB12831) and glycosidases (RB2988, RB2990 and RB2991) were up-regulated at 300 min probably to aid in cell wall remodeling.

Although incorrect protein folding at low temperature is less expected than at high temperatures, chaperons and proteases are required to deal with intracellular protein perturbations (251; 261). Here, this was observed in the induction of GroEL (RB8970) (245; 247) and *htrA*-protease (RB12752). One of the most prominent responses of microorganisms to cold shock is the induction of cold shock proteins. However, the two annotated cold shock proteins of class I (CspA - RB4681 and Cspl - RB10009) (262; 263) were not observed to be regulated. One may hypothesize that the stabilization of RNA in *R. baltica* employs a different protein complement than observed in *E. coli*.

## High salinity

As a marine organism, *R. baltica* must adjust to the haline stratification of the Baltic Sea (264; 265). While moving through the water column *R. baltica* cells are exposed to variable concentrations of dissolved salts. In general, an osmotic up-shift forces bacteria to change their physiology by activating or deactivating specific enzymes or transporters, in order to maintain osmotic balance (266). To gain an understanding of the genetic events that occur during the early stages of salt adaptation, *R. baltica* cells were subjected to salt up-shock from 17.5‰ salinity (Baltic Sea) to 59.5‰ (hyper saline environment). Previous experiments have shown that *R. baltica* is able to grow between salinities of 4.2‰ and 59.5‰ (229) and does not grow at salinities over 90‰ (Wohlrab, unpublished data).

In total, 1127 genes showed differences in gene expression over the whole time series. 656 of these genes (58%) were annotated as hypothetical proteins. The salt up-shock results indicated an increase in the number of regulated genes over time. After 10 min, 61 genes (1%) were regulated. The largest number (543 – 8%) was observed at 300 min (Figure {fig:pirellula1 iii}). *R. baltica* cells seem to adapt slowly to high salt concentration. This might be a result of the cell compartmentalization and resulting ability of *R. baltica* to temporarily resist higher salt concentration without notable cellular responses.

The response of *R. baltica* to salt stress includes repression of genes associated with the following COG classes: induction of amino acid transport and metabolism [E], lipid metabolism [I], transcription [K], translation process [J]. Induced genes were involved in classes of the heat shock experiment (discussed above): [O], [M] and [L]. In addition, genes in the energy production [C] and cell division and chromosome partitioning [D] classes were induced (Figure 10.2 iii). Similar to other bacteria, *R. baltica* accumulated glutamate and trehalose as cytoplasmic osmoprotectants in response to osmotic stress (267). Glutamate dehydrogenase (RB6930) showed an up-regulation after 10 min and was also present in the proteome (247). Trehalose synthetase *treS* (RB519) was induced after 60 min. Cysteine, as a general protective component, was only needed in the first hour in elevated salt concentrations and was repressed afterwards (RB4386).

The accumulation of compatible solutes is a widely distributed mechanism used in coping with changing salinity concentrations (267; 268). In *R. baltica* 74 planctomycetes-group-specific genes are annotated as hypothetical proteins carrying a Domain of Unknown Function (DUF1559) (75). This domain belongs to a new family of solute binding proteins (PF07596) (269) and was also found on the planctomycete fosmid 8FN (75).

Nine of these genes were up-regulated during the first hour of the cold and salt shock experiments. During the heat shock experiment, 16 of these genes were down-regulated. In vitro experiments have shown that some of these compatible solutes also possess general protein stabilization properties in addition to their osmoprotective property (270). These homologous proteins do not play an integral role in the transport process per se, but probably serve as receptors that trigger or initiate translocation of solutes through membranes by binding external sites of the integral membrane proteins of the efflux system. In addition, some solute-binding proteins function in the initiation of sensory transduction pathways (269).

*R. baltica* up-regulated an efflux pump (RB7603) and a  $\text{Na}^+/\text{H}^+$  antiporter (RB1433) 300 min after salt shift. Both may play a role in the active export of salt ions out of the cells. Quinone oxidoreductase-like protein (RB10967), induced after 40 min, had been implicated in respiration-coupled  $\text{Na}^+$  efflux as also shown in *D. vulgaris* (253). Regulatory proteins like sigma-54 factor *rpoN* (RB6491), *rpoA* (RB12626) and *rfaY* (RB12251) were down-regulated. *rpoN* and *rpoA* were found to be translated (245; 247). *R. baltica* inhibited the genes for cell division (*soj* - RB2291) and chromosome segregation (SMC - RB6065) after 60 min salt stress, as well as diverse transferases (RB12080, RB8898, RB12690, RB2498, RB8222, RB9617) involved in the cell envelope modification. Interestingly, the pilin transport apparatus and the thin-pilus basal body (*pilM* - RB2860 and *pilT* - RB12773) were induced after one hour as were principle pilus associated adhesion (*pilC* - RB12781) and *pilB* (RB12774). Genes coding for biopolymer transport proteins (*exbB* - RB12053 and *exbD* - RB12055) were also induced. A homolog to *exbD* was annotated on the planctomycete fosmid 3FN (75). It is known from studies of other organisms that genes encoding the flagellar and chemotaxis systems are up-regulated to move away from the stressful cations (253). However, none of the flagellar genes were regulated and the genome does not harbor any essential chemotaxis genes except *cheY* (21). Notably, the survival protein (SurE - RB10258) and two genes coding for the mechanosensitive ion channel (MscS - RB12279 and RB10255) were induced. The latter provides protection against hypo-osmotic shock, responding both to stretching of the cell membrane and to membrane depolarization (271). Genes in Cluster 22 (ADDITIONAL FILE 2) seemed to be significantly affected by salt stress only.

### 10.2.5 Common stress response

*R. baltica* showed a common stress response to all three tested environmental factors. Several known general stress genes were induced, such as genes coding for the manganese-containing catalase (RB10727), which is also present in the (244; 245; 247). Ferritin and Dps (RB4433) or pyridoxamine 5'-phosphate oxidase (RB4438) belong to a general stress cluster (RB4432 - 4438) and were initially described by Hieu et al. (247). Thioredoxin (RB10378) could serve as an electron donor for the up-regulated methionine sulfoxide reductase gene (*msrB* - RB2268) (272; 273). The genes could be regulated via *rpoN* found on the proposed upstream sigma 54-dependent promoter (RB10378) (274).

Perhaps to cope with reactive oxygen species (ROS), typically present under stressful conditions (273), the nitrogen fixation protein (*nifU* - RB3596) was induced. *nifU* is involved in the biosynthesis and repair of ROS scavenging iron-sulfur clusters. Finally, the peptidase M50 (RB6092) may have been induced to regulate stress response, sporulation, cell division, and cell differentiation (275).

Genes involved in *R. baltica*'s fatty acid metabolism – for example, oA-acyl carrier protein transacylase (*fabD* - RB314), the acyl carrier protein (*acpP* - RB318) and the *fabB* (RB320) gene – were repressed under all conditions.

Interestingly, the machinery for the rearrangement and interchange of genetic material was induced under all three stressful conditions. It seems to play an important role in the organism's long-term adaptation. *R. baltica* harbors 81 non-randomly distributed transposases in its genome. Notably, under heat stress three times more transposase genes were up-regulated than under cold stress and twice as many as under salt stress. Shared

induction shows five IS3/IS911, three ISXo8, two putative transposases (RB170, RB5888, RB11749, RB11802, RB12940, RB2186, RB9907, RB12239, RB934 and RB7389), and one integrase (RB11750). Rearranging the genome to select the most efficient gene combination has been described as a common way to adapt quickly to extreme environments (257). Relaxed DNA may also be required to get better access to the gene regions for increased expression. Here, DNA relaxation is suggested by the repression of histone-like DNA-binding protein (RB6276).

In line with an alternative global sensing and regulation system initially proposed by Glöckner et al. (21), a common pattern concerning sensing and regulation response was detected. *R. baltica* contains 37 genes belonging to the extracytoplasmic function (ECF) subfamily of sigma 70 (276). The genes RB138, RB13241 and RB10049 are up-regulated under all three stress conditions. Studholme et al. (269) suggests that ECF-factor RB10049 is the regulator for the conserved hypothetical protein RB10051. The conserved domain belongs to a new group of proteins that share novel domains referred to as planctomycete-specific (PSD) or planctomycete-specific cytochrome C (PSC). RB10051 contains the PSD1 (DUF1553 - PF07587) and PSC2 (DUF1549 - PF07583) domains, suggesting a function in redox reactions (269). Each domain is represented 41 times in the whole genome of *R. baltica* (75).

Additionally, at 300 min the ECF-sigma factor RB138 was up-regulated together with serine/threonine protein kinase (RB140). Protein kinases are believed to be involved in stress response (260; 277). The serine / threonine protein kinase (RB12942) and two histidine-kinases (RB4511 and RB10330) were up-regulated during heat shock. Whereas, under cold shock only one serine / threonine kinase (RB8505) was induced. Under salt stress a histidine-kinase (RB13122) and three two-component systems (RB5780, RB12952 and RB13118) were induced.

Finally, the ECF-sigma factor RB1790 was up-regulated, but only under high salinity conditions. In summary, the results confirmed that ECF sigma factors, as well as two-component systems, are heavily involved in stress sensing and regulation of *R. baltica*. The importance of these genes in the natural environment is asserted by the presence of a homolog to RB12952 on the planctomycete fosmid 6N14 (75).

The down-regulation of genes associated with the ribosomal machinery (55%) was observed. During heat shock and high salinity these genes were permanently repressed, whereas under cold shock they were only repressed within the first hour. Of the 51 ribosomal proteins in the whole genome, 18 genes encoding proteins of the small- and large subunit (RB1233, RB12821, RB12824, RB12839, RB7117, RB7837-RB7841, RB7849, RB7850, RB7852, RB7854, RB7856, RB7857, RB7859 and RB7899) were repressed. Additionally, a set of genes involved in RNA metabolism, protein synthesis, as well as *R. baltica*'s only translation elongation factor (EF-Tu - RB7894) were repressed. The genes for the conserved hypothetical protein RB12818 and the hypothetical protein RB12837 were co-regulated which suggests an association with the translation machinery. The repression of the ribosomal genes, along with a large set of genes involved in RNA metabolism, protein synthesis, cell growth (Cluster 1 ADDITIONAL FILE 2), has been reported as a general feature of the environmental stress responses (ESR) (256). It has been assumed that they are acting as stress sensors (278). This coincides nicely with the induction of the ribosomal proteins at 300 min under cold shock conditions. Recovery and ongoing adaptation of *R. baltica* was further supported by the up-regulation of the ribosomal-binding factor *rbfA* (RB5503), which is, aside from *csdA*, required for optimal growth at low temperatures (279).

### 10.2.6 Hypothetical proteins

Approximately 50% of the regulated genes observed have no known function in each of the three environmental stress experiments. Some of these share a similar expression profile (ADDITIONAL FILE 2). We propose that some of these genes are involved in cell morphology changes, stress sensing and regulation. The low number of known

transcriptional regulators (2.4%) in the genome of *R. baltica* (276), coupled with the fact that most of the essential pathways encoded are not organized in operon structures (21) support the hypothesis of novel global regulation mechanisms. Hypothetical proteins that carry regulatory domains, like the FHA domain in RB1789 or a putative transcriptional regulatory domain in RB9999 are strong candidates. RB11766 might regulate the gene next to it, which is a so called giant gene (RB11769) (280) This giant gene encodes a novel peptide motif that is most likely involved in cell morphology changes (269). The importance of the hypothetical proteins RB11505, RB10954, RB10956 and RB10958 was further supported by their presence on the proteome gels of Hieu et al. as well as Gade et al. (244; 245; 247). The latter three of these genes were claimed to be among the most abundant proteins in *R. baltica* cultures grown on mineral medium.

### 10.2.7 *Planctomycete* special feature: Genes encoding sulfatases

The genome of *R. baltica* contains no less than 110 sulfatases. It is assumed that they are involved in the recycling of carbon from complex sulfated heteropolysaccharides. Although the mineral medium does not contain any sulfated polysaccharides, we found 11 sulphatase genes were up- or down-regulated (Table 10.3) during the different stress experiments. These included one choline sulphatase (RB1205), seven arylsulfatases (RB13148, RB1477, RB3403, RB406, RB5146, RB684 and RB9498), two sulphatase genes without specificity (RB3956, RB5294), and one alkylsulfatase (RB11502). Furthermore, during life cycle experiments (unpublished data) we found evidence that certain sulfatases are only regulated in specific growth stages, which could indicate their involvement in the remodeling of the distinct morphological features of *R. baltica*. Sulfatase genes RB1477, RB5294, RB9498 and RB11502 were induced. We propose that RB9498 and RB11502 have an extracellular function and may be involved in the formation of an extrapolymeric substance.

Six sulfatase genes (RB406, RB684, RB1205, RB3403, RB5146 and RB13145) were repressed after 300 min of heat shock. They may have been involved in the rearrangement of the cell wall formation, which comprises a protein sacculus with disulfide bonds (237). In summary, these results show the diverse roles that sulfatases may have and, furthermore, that only a variety of different experimental approaches will increase our knowledge of these roles.

## 10.3 Conclusion

This work presents the first transcriptome study of the environmental stress response of a marine, free-living *Planctomycete*. Although *R. baltica* is an unusual organism in many aspects, its stress responses to heat and cold shock as well to changing salinity were in line with earlier results reported for other model organisms. Heat shock induced a set of chaperons, likely to protect cellular proteins from denaturation and breakdown. Growth in the cold may be followed by the induction of genes altering lipid metabolism. Salinity shifts resulted in the activation of planctomycete-specific groups of genes including genes involved in morphological change and an extracytoplasmic stress response. All stressors triggered the down-regulation of the ribosomal machinery, the up-regulation of transposases and the induction of several ECF-sigma factors and two-component systems. This supports the hypothesis that *R. baltica* is regulating its gene activity on a global rather than operon level. Aside from well characterized stress response genes, about 2000 genes of unknown function, constituting 30% of the genes predicted in the genome, were affected. This, combined with proteome studies and the presence of some of the genes in fosmid libraries, provides a strong indication that the vast number of genes with unknown function play a vital role in the organism's environmental response. The regulation of 11 sulfatases during stressful conditions suggests that these genes are heavily involved in the core cellular function of *R. baltica*. The data presented lead to the conclusion that

*R. baltica*'s rich repertoire of genes is combined with a fine tuned regulation mechanism to best respond to the changing conditions of its habitat. Nevertheless, data analysis has just started and further investigations concerning the genes involved in the life-cycle, the stress response pathways, promoter regions and network analysis are already ongoing or planned for the near future.

## 10.4 Methods

### 10.4.1 Bacterial growth conditions

For all experiments *Rhodopirellula baltica* SH1<sup>T</sup> cells were grown as chemostat cultures in a mineral medium containing 10 mM glucose as the sole carbon source and 1 mM ammonium chloride as a nitrogen source at 28°C (243). Chemostat (Ø 13.5 cm x 25 cm, 1 l, Schott, modified by Ochs, Bovenden) parameters used were: pH 7.4, average dilution rate 0.75 ml/min and pO<sub>2</sub> around 100%. The cultures had an OD<sub>600nm</sub> of 0.5 – 0.6 (corresponding to log phase). The cells were harvested after 5 dwell times.

### 10.4.2 Sample collection, cell lyses, RNA Isolation and cDNA synthesis

After harvesting the *R. baltica* cultures, an aliquot was collected to serve as the time-zero reference. The culture broth was collected in 500 ml tubes and swirled briefly in an ethanol-dry ice bath to rapidly cool the cultures and prevent shifts in the RNA profile. Subsequently, the broth was centrifuged at 6000 rpm for 20 min at 4°C (Beckman Coulter™ Avanti™626 J-20XP, JA10 Rotor). The pellets were re-suspended in 0.1 M Tris-HCL and then re-centrifuged. Cell pellets were shock-frozen in liquid nitrogen and stored at -80°C. Total RNA was isolated using the protocol of the TRI Reagent® Kit by Ambion(Austin, USA). The purity and quality of the extracted total RNA was checked with an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, USA) and gel electrophoresis. cDNA synthesis was performed using the SuperScript direct cDNA labeling kit by Invitrogen (Karlsruhe, Germany) according to the manufacturer's instructions with random hexamers and unlabeled dCTP/dUTP, followed by a three hour reverse transcription incubation step at 46°C. The RT reaction was halted by incubation for 3 min at 95°C. To hydrolyze the RNA, 0.1 M NaOH was added, incubated at 65°C for 15 min and neutralized with 0.1 M HCL. The remaining cDNA was precipitated overnight at -20°C and the pellet washed with 70% Ethanol.

cDNA was directly labeled using the PlatinumBright™ nucleic acid labeling kit based on KREATECH's patented Universal Linkage System (ULS) (Biocat, Heidelberg, Germany) according to the manufacturer's protocol.

Concentrations of RNA and cDNA were measured, and incorporation of the dyes Alexa 546 and Alexa 647 were checked using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, USA).

### 10.4.3 Experimental design and sample preparation

In three independent hybridizations conducted for each experiment and time point, the expression profiles of cells that had undergone stress were compared with those of cells at time zero. That is, the array analysis of each Alexa 647 labeled sample was compared with those of Alexa 546 labeled time-zero samples. The data shown are based on the analysis of all three replicates performed for each of the conditions.

Samples for expression profiling and microscopic analysis were collected at 10, 20, 40, 60 and 300 min in all three stress experiments.



#### 10.4.4 Heat shock from 28°C to 37°C

Cells grown continuously at 28°C were collected by centrifugation. An aliquot was removed for RNA extraction and taken as the time zero reference for the heat, cold and salt stress experiments. Aliquots were re-suspended in an equal volume of 37°C medium and returned to 37°C for cultivation.

#### 10.4.5 Cold shock from 28°C to 6°C

Cells grown continuously at 28°C were collected by centrifugation, re-suspended in an equal volume of 6°C medium and returned to 6°C for cultivation.

#### 10.4.6 Salt stress from 17.5‰ to 59.5‰ salinity

Similar to the heat and cold shock experiments, an *R. baltica* culture was grown in mineral media with 17.5‰ salinity. Cells were harvested and aliquots were transferred to a mineral media with a salinity of 59.5‰.

#### 10.4.7 Whole Genome Array construction, hybridization and image analysis

The whole-genome oligonucleotides for *R. baltica* SH1<sup>T</sup> (*Pirellula* AROS 630 Version 1.0) were purchased from Operon (Cologne, Germany) and diluted to 20  $\mu$ M concentration in Micro Spotting Solution Plus spotting buffer (Telechem, Sunnyvale, USA). Spotting was done with three replicates per gene, per slide onto GAPS II aminosilane slides (Corning, Schiphol-Rijk, Netherlands) using a SpotArray 24 spotting device (Perkin Elmer, Wellesley, USA) together with 48 Telechem Stealth Pins (Telechem, Sunnyvale, USA). The arrays were subsequently exposed at 245 nm and 360 mJ in the GS Gene Linker (Bio-Rad, München, Germany), followed by incubation at 80°C for at least 3 h. Slides were stored at room temperature in the dark until use.

Blocking, denaturing, hybridization, washing and N<sub>2</sub> drying procedures were carried out in an automated hybridization station HS400 (Tecan, Crailsheim, Germany). The spotted arrays were blocked in prehybridization solution containing 250 mM NaCl, 5 mM Tris/HCl at pH 8.0, 50% formamide, 0.5x SSC, 0.05% BSA, and 1% blocking reagent from Roche Diagnostics, Mannheim, Germany for 45 min at 52°C. For hybridization at least 2  $\mu$ g of Alexa 546 dye-labeled and 2  $\mu$ g of Alexa 647 dye-labeled total cDNA were combined and taken up in a final volume of 100  $\mu$ l DIG Easy Hyb hybridization solution (Roche Diagnostics, Mannheim, Germany). After the blocking step, the sample solution was applied to the arrays, denatured at 95°C for 3 min and hybridized under stringent conditions at 52°C for over 12 hours. After hybridization slides were washed at room temperature in ULTRArray Low Stringency Wash Buffer (Ambion, Austin, USA) and dried by N<sub>2</sub>.

#### 10.4.8 Signal detection and data analysis

Slides were scanned at a resolution of 5  $\mu$ m using a ScanArray Express Microarray scanner (Perkin Elmer, Wellesley, USA) with varied laser power and photomultiplier tube (PMT sensitivity) for each slide. The accompanying image analysis software, ScanArray Express Version 4.0, was used for automatic spot detection and signal quantification of both fluorophores. Raw data were automatically processed using the microarray data analysis software tool MADA ([www.megx.net/mada](http://www.megx.net/mada)), developed in-house. Firstly, the spot intensities were corrected for local background (mean spot intensity minus mean spot background intensity). Signals were only assessed as positive if mean spot pixel intensity was higher than the mean local background intensity plus twice the standard deviation of the mean local background pixel intensity. Each gene is spotted in three replicates. Spot replicates with poor quality were removed from the data set according

to MADA's outlier test results. This test first computes the standard deviation of all replicates. Secondly, one replicate is omitted and the standard deviation is recalculated; if the deviation differs more than 50% from the previous deviation, the omitted replicate is regarded as an outlier. This procedure is repeated for all replicates

Expression is described through the ratio and intensity, where R is the fluorescence log ratio of the experiment time point relative to the control condition (e.g.  $R = \log_2(\text{result of channel 10min} / \text{result of channel control/reference})$ ) and I is the log mean fluorescence intensity (e.g.  $I = \log_{10}(\text{result of channel 10 min} \times \text{result of channel control} / \text{reference})$ ).

Each data point represents a regulation factor (ratio) in a logarithmic scale for one gene calculated from the positive replicates for a particular probe coming from two RNA pools (reference and sample). Normalization was carried out by LOWESS fitting on an R-versus-I plot with a smoothing factor of 0.5. Each time point of the time-series experiment was hybridized independently three times. The expression data (ratio) of the three hybridizations were combined to one expression data point (ratio) by averaging and the standard deviation of the average value was calculated. Only ratios with a standard deviation less than 25% were regarded as genes that are regulated. Differentially expressed genes were detected by a fixed threshold cut off method (i.e. a two-fold increase or decrease) based on the results of self-self hybridization. Using the same biological sample, the reference (untreated sample) is labeled twice, once with Alexa 546 and once with Alexa 647, and the variability between the two sets of measurements is calculated to estimate the experimental noise. Ideally, there should not be any variability and all expression points should have a ratio close to zero. In reality, however, this is never the case and thresholds based on the distribution of these data along the y-axis were defined for the further experiments.

Consequently, *R. baltica* genes detected with intensities resulting in ratios above or below these thresholds can be regarded as up- or down-regulated.

#### 10.4.9 Cluster analysis

Differentially expressed genes present in the complete time course profile (10, 20, 40, 60 und 300 min) for all three experiments were clustered using the k-means clustering approach (Euclidean distance metric,  $k = 30$  clusters and 49 (max. 500) iterations) (281) with the software tool Multiexperiment Viewer MeV Version 4.0.2 from the TM4 microarray software suite (282). Briefly, the clustering algorithm arranges genes into a given number of clusters,  $k$ , according to similarity in their expression profiles across the entire array experiments, such that genes with similar expression patterns are clustered together. The data are displayed in tabular format where each row of colored boxes stores the variation in transcript abundance for each given gene and each column stores the variation in transcript levels of every gene in a given mRNA sample, as detected on one array. The variations in transcript abundance for each gene are depicted by means of a color scale, in which shades of red represents increases and shades of green represent decrease in mRNA levels, relative to the unstressed culture, and the saturation of the color corresponds to the magnitude of the differences. Black coloration indicates no change in transcript level while grey represents missing data.

#### 10.4.10 Genome tools

The genome of *Rhodopirellula baltica* was automatically re-annotated based on updated homology searches (June 2005 - MicHanThi (283)). The updated annotation including all tool results are publicly available at <http://gendb.mpi-bremen.de/gendb/BX119912> (284). JCoast (285) was used as a tool for the visualization, interpretation, COG-assignment statistics and comparison of genomic data stored in GenDB V2.2 (9). The Venn diagrams were generated by BioVenn (286).

### 10.4.11 Microarray Datasets

Each microarray used in this study contained 7325 known or predicted *R. baltica* genes according to Glöckner et al. (21). A detailed description of the array can be found at the NCBI's Gene Expression Omnibus (GEO) database under accession number GPL7654. The complete microarray datasets covering the expression of *R. baltica* cultures exposed to heat, cold and high salinity, are public available in the GEO repository (<http://www.ncbi.nlm.nih.gov/geo/>) under accession numbers GSE13769, GSE13856 and GSE14075 (287).

## 10.5 List of abbreviations

**COG:** Cluster of Orthologous Group of Genes.

**DUF:** Domain of Unknown Function.

**ECF:** Extra Cytoplasmic Function.

**ESR:** Environmental Stress Response.

**FHA:** Forkhead-associated.

**GEO:** Gene Expression Omnibus.

**R.:** *Rhodopirellula*.

**RB:** *Rhodopirellula baltica*.

**ROS:** Reactive Oxygen Species.

**ORF:** Open Reading Frame.

## 10.6 Competing interests

The authors declare that they have no competing interests.

## 10.7 Authors contribution

PW conceived the study, initiated, conducted the experimental analysis, validated microarray and optimized experimental steps, wrote the manuscript, did the statistical analysis and analyzed the data. CK was involved in the chemostat cultivation of *R. baltica* and statistical analysis and in the biological interpretation of the data. AE wrote MADA the microarray analysis tool and designed the microarrays. CQ was responsible for the automatic reannotation of the genome and set up the web access PL and JH established the chemostat cultivation of *R. baltica*. JH supervised the chemostat cultivation. FOG contributed background information and was involved in writing and finishing the manuscript. All authors read and approved the final manuscript.

## 10.8 Acknowledgements

We thank Sylke Wohlrab for the preliminary work leading to this study, Matthias Gottschall for excellent technical assistance in the chemostat cultivation, and Pier Luigi Buttigieg for language revision.

The setup of the microarray and transcriptomic analysis of *Rhodopirellula baltica* was supported by the Federal Ministry of Education and Research contract 03F0364A and the European Commission contract GOCE-CT-2004-505403.

**Table 10.1:** *Shared stress response to heat, cold and high salinity: Results for induced genes are shown.*

ID	AA	Product	IEP	Strand	Potentially involved in / Comments
RB170	96	Transposase IS3/IS911	10.1	+	
RB370	553	nitrate transporter substrate-binding protein	4.6	+	
RB521	63	hypothetical protein	10.7	-	(247; 248)
RB723	60	hypothetical protein	7.3	+	
RB934	375	Putative transposase	9.9	-	
RB1394	78	hypothetical protein	10.1	-	regulatory mechanism
RB1395	319	secreted protein similar to DNA-binding protein	5.5	+	regulatory mechanism
RB1789	243	conserved hypothetical protein	9.4	+	regulatory mechanism
RB1872	38	hypothetical protein	12.3	+	
RB2186	433	ISXo8 transposase	9.4	-	
RB2268	282	peptide methionine sulfoxide reductase	9.9	-	(245)
RB3596	144	nitrogen fixation protein (NifU protein)	4.3	+	
RB4299	96	Transposase IS3/IS911	9.9	+	
RB4347	156	conserved hypothetical protein	4.7	+	(247; 248)
RB4397	55	protein containing DUF1560	10.5	+	
RB4429	89	conserved hypothetical protein	4.9	+	stress response
RB4433	162	Ferritin and Dps	4.3	+	
RB4438	160	Pyridoxamine 5'-phosphate oxidase-	4.4	+	
RB4510	49	hypothetical protein	9	-	
RB5238	73	hypothetical protein	10.6	+	
RB5551	663	hypothetical protein	5.7	+	DVL-domain
RB5888	96	Transposase IS3/IS911	10.1	-	
RB5938	370	hypothetical protein-	5.6	-	(247)
RB6928	160	hypothetical protein	4.1	+	
RB7389	375	Putative transposase	9.9	+	
RB8409	97	hypothetical protein	8.9	+	
RB8527	330	protein containing DUF1559	6.6	+	stress response
RB8987	48	hypothetical protein	9.1	-	
RB9230	107	hypothetical protein	9.9	+	next to transposase
RB9907	433	ISXo8 transposase	9.4	+	

Continued on next page

Table 10.1 – continued from previous page

ID	AA	Product	IEP	Strand	Potentially involved in / Comments
RB9955	452	secreted protein containing DUF1552	5.6	-	regulatory mechanism
RB9999	281	conserved hypothetical protein-	4.6	-	regulatory mechanism
RB10049	217	RNA polymerase ECF-type sigma factor	10.1	+	
RB10378	144	Thioredoxin	4.6	-	
RB10727	276	manganese-containing catalase	5	+	(247; 248)
RB10728	132	secreted protein	9.9	+	stress response
RB10896	161	secreted protein	10	-	stress response
RB10954	143	hypothetical protein	10.4	-	
RB10956	117	hypothetical protein	4.8	+	(248)
RB10957	99	conserved hypothetical protein	5.6	+	regulatory mechanism
RB10958	158	hypothetical protein	5.4	+	
RB11176	153	protein containing DUF442	4.8	-	(248) stress response
RB11260	121	dnaK suppressor protein,	5.2	-	
RB11475	57	conserved hypothetical protein	4.7	+	next acyltransferase, short protein
RB11504	72	conserved hypothetical protein	10.7	-	short protein,
RB11505	199	conserved hypothetical protein, secreted	7.5	-	
RB11515	74	conserved hypothetical protein	11.7	+	
RB11566	195	hypothetical protein	10.8	+	
RB11749	96	Transposase IS3/IS911	10.1	+	
RB11750	292	integrase	10	+	
RB11802	96	Transposase IS3/IS911	10.1	+	
RB11855	101	conserved hypothetical protein	12.5	-	
RB11918	134	protein containing DUF971	6.1	-	
RB11977	196	conserved hypothetical protein	9.5	+	
RB12066	135	hypothetical protein	10	+	
RB12239	433	ISXo8 transposase	9.4	+	
RB12247	74	conserved hypothetical protein	6.3	-	
RB12936	580	conserved hypothetical protein	5.3	-	DUF 444
RB12940	96	Transposase IS3/IS911	10.1	+	
RB13222	208	SOUL heme-binding protein	8.7	-	
RB13241	167	RNA polymerase ECF-type sigma factor	8.8	-	

**Table 10.2:** *Shared stress response to heat, cold and high salinity: Results for repressed genes shown.*

ID	AA	Product	IEP	Strand	Potentially involved in / Comments
RB61	58	hypothetical protein	9.6	+	
RB314	309	malonyl CoA-acyl carrier protein transacylase	4.4	+	
RB318	81	Acyl carrier protein	3.7	+	
RB319	95	hypothetical protein	10.2	+	fatty acid process
RB767	311	conserved hypothetical protein, secreted	5.7	+	
RB825	117	hypothetical protein	7.8	-	
RB951	234	protein containing DUF1596	12.3	-	
RB1129	895	conserved hypothetical protein	5.8	-	(S) bombinin, defense response
RB1233	206	30S ribosomal protein S4	11.2	-	
RB2105	470	membrane protein	9.6	-	
RB2306	41	hypothetical protein	9	+	
RB2479	273	conserved hypothetical protein	5.5	-	
RB3277	221	hypothetical protein	10.4	-	
RB3362	87	hypothetical protein	11.9	-	
RB3366	78	hypothetical protein	5.6	+	
RB3394	36	hypothetical protein	10.5	+	
RB3399	65	hypothetical protein	9.7	-	
RB3575	152	membrane protein	10.1	+	(244)
RB3603	344	secreted protein	4.6	-	
RB3675	742	secreted protein	8.4	+	
RB3688	53	hypothetical protein	9.3	-	
RB3880	82	hypothetical protein	10.7	-	
RB3953	857	hypothetical protein	5.2	-	
RB3981	161	hypothetical protein	4.1	+	
RB3994	191	hypothetical protein	4.1	+	
RB4097	733	conserved hypothetical protein	6.2	+	
RB4145	90	hypothetical protein	12	-	
RB4194	53	hypothetical protein	11.4	-	next to a seromine/threonine kinase
RB4269	282	glutamic acid specific endopeptidase	5.6	-	
RB4358	123	hypothetical protein	6.5	-	(247; 248)

Continued on next page

Table 10.2 – continued from previous page

ID	AA	Product	IEP	Strand	Potentially involved in /	Comments
RB4373	109	hypothetical protein	4.8	-		
RB4657	123	hypothetical protein	12.2	+		
RB4951	95	hypothetical protein	12.1	+		
RB5262	95	membrane protein	6.3	+		
RB5409	97	hypothetical protein	12.7	+		
RB5415	62	hypothetical protein	12.3	+		
RB5745	130	hypothetical protein	10.7	-		genetic information processing
RB6092	361	Peptidase M50	9.6	+		
RB6158	142	hypothetical protein	6	-		
RB6174	69	hypothetical protein	10.6	-		
RB6276	105	Histone-like bacterial DNA-binding protein	10.4	-		
RB6634	365	protein containing DUF1559	5.3	+		
RB6699	47	hypothetical protein	11.1	+		
RB6766	55	hypothetical protein	12	+		
RB6849	101	hypothetical protein	12.8	-		
RB7042	91	hypothetical protein	10.4	+		
RB7116	59	hypothetical protein	11.7	+		ribosomal machinery
RB7117	181	Ribosomal protein L35	11.4	-		
RB7557	327	von Willebrand factor type A domain protein	4.9	+		
RB7646	62	hypothetical protein	10.5	+		
RB7647	73	hypothetical protein	7.4	+		
RB7837	286	Ribosomal protein L2	11.8	+		
RB7838	89	Ribosomal protein S19/S15	10.8	+		
RB7839	119	Ribosomal protein L22/L17	11	+		
RB7840	236	30S ribosomal protein S3	10.4	+		
RB7841	138	Ribosomal protein L16	11.1	+		
RB7849	108	Ribosomal protein S17	10	+		
RB7850	122	Ribosomal protein L14b/L23e	11	+		
RB7852	196	50S ribosomal protein L5	10.4	+		
RB7854	61	Ribosomal protein S14	11.8	+		
RB7856	181	50S ribosomal protein L6	10	+		
RB7857	149	Ribosomal protein L18P/L5E	11.6	+		

Continued on next page

Table 10.2 – continued from previous page

ID	AA	Product	IEP	Strand	Potentially involved in / Comments
RB7859	177	Ribosomal protein S5	10.6	+	
RB7894	398	translation elongation factor EF-Tu	5.2	+	(247; 248)
RB7899	141	50S ribosomal protein L11	9.6	+	
RB8119	142	hypothetical protein	10.1	+	
RB8457	113	hypothetical protein	11.5	-	
RB8594	41	hypothetical protein	9.2	+	
RB8669	37	hypothetical protein	11.5	+	ribosomal machinery
RB9343	59	hypothetical protein	11.4	+	
RB9417	103	hypothetical protein	10.5	+	
RB9460	79	hypothetical protein	10.3	-	
RB9872	67	hypothetical protein	5.4	+	cell division related
RB10581	384	secreted protein containing DUF1559	6.2	+	(248)
RB11287	75	hypothetical protein	9.1	+	
RB11392	148	conserved hypothetical protein	5	+	
RB11490	181	conserved hypothetical protein, membrane	10.3	+	
RB11707	83	conserved hypothetical protein	9.8	+	stress function
RB11766	129	hypothetical protein	10.4	+	(269)
RB12193	36	hypothetical protein	7.5	-	overlapping with asnB (RB12191)
RB12251	567	RNA polymerase specialized sigma factor	9.4	-	
RB12327	686	TGF-beta receptor, type I/II extracellular region	4.5	-	
RB12329	110	conserved hypothetical protein, membrane	4	+	
RB12396	57	hypothetical protein	11.3	-	
RB12454	199	hypothetical protein	10.3	-	
RB12818	163	conserved hypothetical protein	11.7	-	ribosomal machinery
RB12821	117	Ribosomal protein L19	11.1	-	
RB12824	146	Ribosomal protein S16	5.3	-	
RB12837	65	hypothetical protein	9.8	+	ribosomal machinery
RB12839	225	Ribosomal protein L1	9.8	+	



Table 10.3: Differentially expressed sulfatase genes of *R. baltica* are shown.

ID	Product	AA	SignalP	Heat	Cold	Salt	Remarks
RB1205	choline sulfatase	456	0.80	repressed	repressed		
RB3403	arylsulfatase precursor	491	0.99	repressed	repressed		[23, 25]
RB3956	sulfatase	489	0.98			repressed	
RB5146	arylsulfatase A precursor (ASA)	522	0.95	repressed			
RB9498	arylsulfatase A	518	0.97	induced			[23,25]
RB11502	alkyl sulfatase or beta-lactamase	445	1.00			induced	
RB1477	arylsulfatase precursor	538	0	induced		induced	
RB5294	sulfatase	533	0			induced	wall* unpublished results
RB406	arylsulfatase	557	0	repressed	repressed		
RB684	arylsulfatase precursor	653	0	repressed		repressed	life cycle unpublished results
RB13148	arylsulfatase A [precursor]	1012	/	repressed			life cycle unpublished results



# Chapter 11

## Megx.net Paper

### Megx.net – database resources for marine ecological genomics

T. Lombardot<sup>a</sup>, R. Kottmann<sup>a</sup>, H. Pfeffer<sup>a</sup>,  
M. Richter<sup>a</sup>, H. Teeling<sup>a</sup>, C. Quast<sup>a</sup>  
and F.O. Glöckner<sup>a,b</sup>

<sup>a</sup>Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany;

<sup>b</sup>Jacobs University Bremen gGmbH, D-28759 Bremen, Germany;

**Journal; Volume (Issue):** Nucl. Acids Res.; 34

**Pages:** D390-393

**Month / Year:** 2006

**DOI:** 10.1093/nar/gkj070

**Contributions:**

Technical support for the setup of the web server.

---

**Contents**

---

<b>11.1 Introduction</b>	<b>143</b>
<b>11.2 Sources of Genomic and Metagenomic Data</b>	<b>144</b>
<b>11.3 Genome Browsing</b>	<b>144</b>
<b>11.4 Precomputed Information</b>	<b>144</b>
11.4.1 Environmentally relevant protein families	144
11.4.2 Group-specific genes	145
<b>11.5 TETRA Server</b>	<b>145</b>
<b>11.6 Genomes Mapserver</b>	<b>145</b>
<b>11.7 Additional Features</b>	<b>146</b>
<b>11.8 Databases Access</b>	<b>146</b>
<b>11.9 Acknowledgements</b>	<b>146</b>

---

## Abstract

Marine microbial genomics and metagenomics is an emerging field in environmental research. Since the completion of the first marine bacterial genome in 2003, the number of fully sequenced marine bacteria has grown rapidly. Concurrently, marine metagenomics studies are performed on a regular basis, and the resulting number of sequences is growing exponentially. To address environmentally relevant questions like organismal adaptations to oceanic provinces and regional differences in the microbial cycling of nutrients, it is necessary to couple sequence data with geographical information and supplement them with contextual information like physical, chemical and biological data. Therefore, new specialized databases are needed to organize and standardize data storage as well as centralize data access and interpretation. We introduce Megx.net, a set of databases and tools that handle genomic and metagenomic sequences in their environmental contexts. Megx.net includes (i) a geographic information system to systematically store and analyse marine genomic and metagenomic data in conjunction with contextual information; (ii) an environmental genome browser with fast search functionalities; (iii) a database with precomputed analyses for selected complete genomes; and (iv) a database and tool to classify metagenomic fragments based on oligonucleotide signatures. These integrative databases and webserver will help researchers to generate a better understanding of the functioning of marine ecosystems. All resources are freely accessible at <http://www.megx.net>.

## 11.1 Introduction

Over the last decade microbiology has undergone several changes. Robert Koch's invention of pure culture techniques at the end of the 19th century focussed microbiology on the isolation of bacteria for laboratory studies. In 1987 Carl Woese introduced the ribosomal RNA as a stable molecular marker for the classification and identification of microorganisms (288). The 'winds of change' blew in the field of microbiology (289) when the first cultivation-independent investigations reported an immense array of completely unexpected microbial diversity in the environment (16). The landmark publication of the first complete genome sequence of *Haemophilus influenzae* in 1995 (14) has transformed biology into a massively parallel and high throughput endeavour. This 'genomic revolution' finally reached the field of marine ecological genomics in the year 2000, defined as: 'The application of genomic sciences to understanding the structure and function of marine ecosystems' (290). Since 1995, >260 microbial genomes have been fully sequenced, and 600 more are well on their way (290). While most projects focus on microorganisms of medical or biotechnological interest, 22 complete marine genomes of environmental organisms are already available, and 130 marine isolates are currently sequenced (Moore foundation <http://www.moore.org>). Recently, this cultivation-based approach has been complemented by a number of groundbreaking cultivation-independent—metagenomic—studies, the most prominent being the Venter Sargasso Sea expedition in 2004 (22), delivering >1.2 million new genes. This wealth of information caused a quantum leap in marine sciences and demands for different kinds of databases to transfer information into knowledge (291). The sequences, genomes, genes and predicted metabolic functions can not longer be regarded in an organism centric view but have to be handled in the context of the environment surrounding them. Therefore, it is necessary to link any environmental sequence information with its geographical location. This allows to correlate the genomic features found at a distinct sampling site with physical, chemical and biotic information to identify organism-specific adaptations and their role and impact on the environment. This new kind of integrative data resource opens the path to address questions like: Are there differences in the genetic repertoire when travelling from coastal marine sites to the open ocean? or Do habitat specific gene patterns with yet unknown functions exist? If the latter is true the correlation with site specific environmental parameters might allow predicting a potential function for them. Can these genetic properties in turn explain the distribution of the organisms?

Megx.net is designed to tackle these tasks linking marine genome and metagenome sequences not only with geography but providing additional information about annotation highlights, presence of environmentally relevant protein families and group-specific genes as well as a Geographic-BLAST server to trace genes across the marine environment.

## 11.2 Sources of Genomic and Metagenomic Data

The genome sequences of all currently available marine microorganisms have been retrieved from the EMBL and GenBank databases (67; 292). Twenty-two bacteria and archaea originating from the water column of the ocean and from marine sediments have been completely sequenced (October 2005). The sequences and associated gene annotation have been imported into a local relational database allowing fast data retrieval. The corresponding annotations originate from independent submissions to the EMBL or the GenBank databases, and are of variable quality owing to the following reasons: (i) the original annotations were performed at different times; (ii) no controlled vocabulary is used for gene product names; and (iii) the effort expended in assigning functions to genes is variable between genome projects. Ecologically relevant annotation highlights were selected from original genome publications for each organism.

Metagenomic fragments originating from marine systems have been selected according to semi-automatic literature screening. Seventy-eight original publications were found to deal with metagenomic fragment sequencing, corresponding to a total of 21 distinct marine geographic sampling sites (August 2005). The sequences and associated gene annotation were imported into a newly designed geographic database. New genomes or metagenomes will be integrated in the database and mapserver as soon as they become available. Precomputed searches will be updated every 2 months.

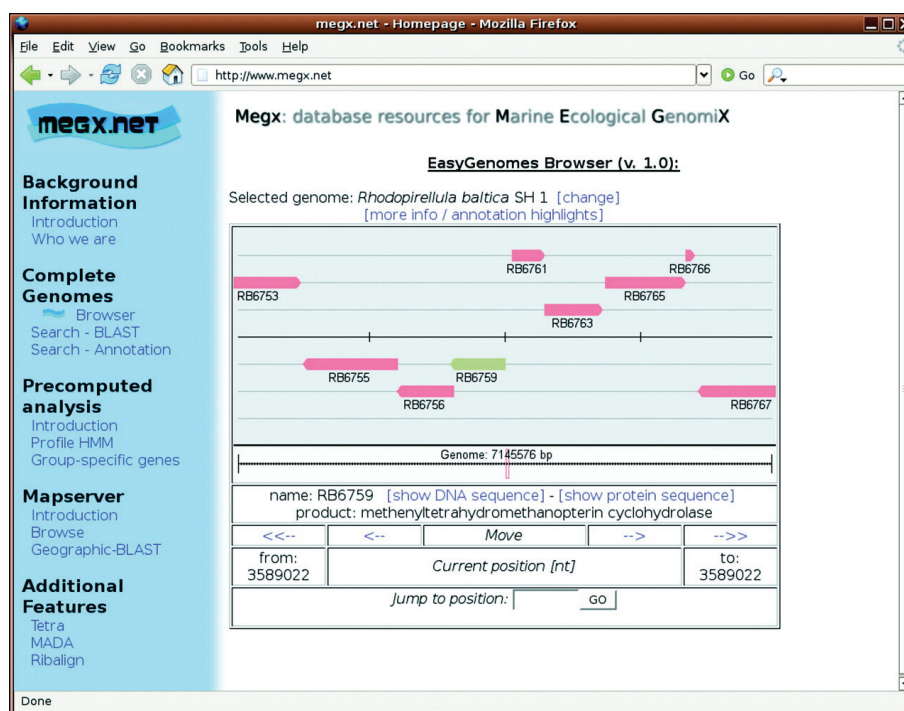
## 11.3 Genome Browsing

The genome browser allows easy and fast access to the sequences, their geographical location and the annotation highlights of each marine microorganism in the database. For example, the unexpected archaea-like C1 metabolism genes found in the genome of *Rhodospirellula baltica* can be accessed in their genomic context by a simple mouse click (Figure 11.1). Fast text search in the original annotations and BLAST searches are also available.

## 11.4 Precomputed Information

### 11.4.1 Environmentally relevant protein families

Some gene families are of particular interest for ecological genomics, as they play key roles in the environment or give insights into the adaptation of microorganisms to their respective niche. Glycosylhydrolases, sulphatases, peptidases and transcriptional regulators are some examples of gene groups that have been automatically extracted based on selected profile hidden Markov models originating from the Pfam database (293). The results can be browsed graphically on our web page. This search strategy allows consistent quantitative comparisons, as the publicly available original annotation can not easily be compared. For example, the outstanding number of genes encoding sulphatases in the genome of *R. baltica* (21) or the reduced dataset of transcriptional regulators in *Prochlorococcus marinus* strains (294; 295) can be compared with the corresponding gene content of other marine microorganisms.



**Figure 11.1:** Fast access to the annotation highlights of marine microorganisms. Here, the archaea-like C1 metabolism key gene is *R.baltica*.

### 11.4.2 Group-specific genes

Group-specific genes are defined as those found exclusively in a defined subset of genomes. The definition of groups is variable and can be based on a phylogenetic affiliation, a common metabolism or related habitats. An example for group specific genes for phylogenetically closely related organisms are the three available *P.marinus* strains. The results show that some light-inducible proteins are exclusively found in those organisms (295). Moreover, we present a set of proteins of yet unknown function which are *P.marinus* specific. The corresponding genes represent interesting targets for functional genomics and further wet-lab experiments.

## 11.5 TETRA Server

TETRA is a software tool for genomic and metagenomic analysis. It can assess the relatedness of genomic fragments by computing correlations between their tetranucleotide usage patterns (i.e. statistical over- and under-representation of tetranucleotides) (296; 181). The new version includes chaos game plot representations for DNA sequences, which can be used to get additional information on the relatedness of genomic fragments. Moreover, TETRA can plot fluctuations of tetranucleotide usage patterns within DNA sequences. This is particularly useful to identify irregular regions in entire genomes or larger genomic fragments like laterally transferred genes or transposase and phage insertions.

## 11.6 Genomes Mapservers

Geographic information systems (GIS) are commonly used in the field of geology for data integration. A GIS is a combination of elements designed to store, retrieve, analyse and display geographic data. We introduce here the Genomes Mapservers, a GIS that allows

access to genomic and metagenomic sequence data in their geographic and ecological contexts. The sampling sites of marine (meta)-genomic studies are displayed within a browsable world map (Figure 11.2). Each sampling site can be selected to display the corresponding sequences and additional contextual information. The underlying database is designed to enable future data mining tasks to reveal possible gene patterns associated with a particular environmental context. For targeted searches, a geographic-BLAST tool has been developed, allowing to perform 'spatial' queries for sequences based on the popular BLAST algorithm (31) The Geographic-BLAST/Genomes Mapservers combination allows to systematically study the biogeography of particular genes in the environment (Figure 11.2).

## 11.7 Additional Features

A software tool for microarray data evaluation and a database of aligned ribosomal proteins for phylogenetic analysis (Ribalign) will soon be available on the webpage.

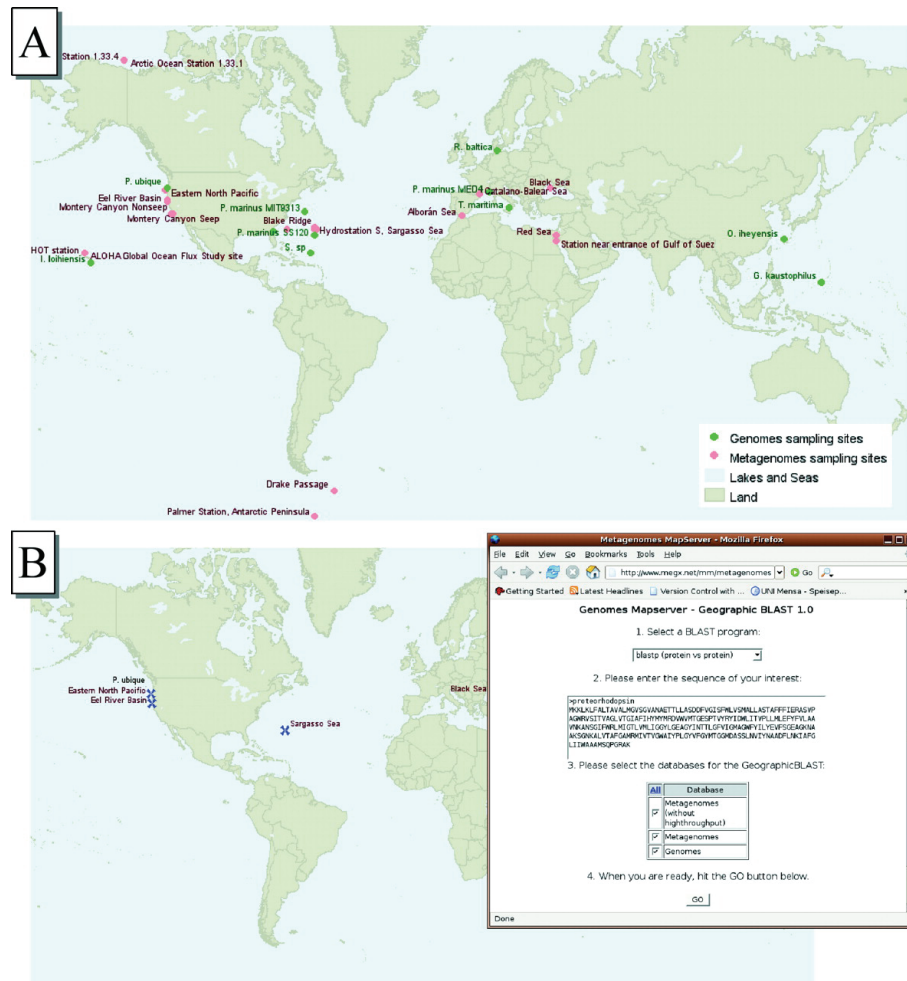
## 11.8 Databases Access

The precomputed genome searches and group-specific genes, the TETRA server and the Metagenomes Mapservers are freely available through <http://www.megx.net>.

## 11.9 Acknowledgements

We thank the Max Planck Society for initial funding and the EU Sixth Framework Programme (FP6-NEST) for providing financial support for further development of the Genomes Mapservers (contract no. 511784). Funding to pay the Open Access publication charges for this article was provided by the Max Planck Society.





**Figure 11.2:** *The Genomes Mapserver. (A) Marine genomes and metagenomic fragments can be browsed and searched on a world map on our web-based system. (B) An example showing a Geographic-BLAST search for genes encoding proteorhodopsins in the currently available dataset.*



**Part III**  
**Appendix**



# Appendix A

## Tools, Libraries & Databases

**Table A.1:** *Build utilities, programming libraries, bioinformatics tools, databases, and third party resources used in this thesis.*

Name	Used In	URL
<b>Build Utilities:</b>		
Apache Ant	MicHanThi	<a href="http://ant.apache.org/">http://ant.apache.org/</a>
Autoconf	SILVA	<a href="http://www.gnu.org/software/autoconf/">http://www.gnu.org/software/autoconf/</a>
Automake	SILVA	<a href="http://www.gnu.org/software/automake/">http://www.gnu.org/software/automake/</a>
libtool	SILVA	<a href="http://www.gnu.org/software/libtool/">http://www.gnu.org/software/libtool/</a>
<b>Programming Libraries:</b>		
ARB	SILVA	<a href="http://www.arb-home/">http://www.arb-home/</a>
Boost	SILVA	<a href="http://www.boost.org/">http://www.boost.org/</a>
JSAP	MicHanThi	<a href="http://www.martiansoftware.com/jsap/">http://www.martiansoftware.com/jsap/</a>
libbz2	SILVA	<a href="http://www.bzip.org/">http://www.bzip.org/</a>
libmysqlclient	SILVA	<a href="http://www.mysql.com/">http://www.mysql.com/</a>
libpcre / libpcrecpp	SILVA	<a href="http://www.pcre.org/">http://www.pcre.org/</a>
libphoenix	SILVA	<a href="http://www.bioinformatics.org/phoenix/wiki/">http://www.bioinformatics.org/phoenix/wiki/</a>
libz	SILVA	<a href="http://www.zlib.net/">http://www.zlib.net/</a>
mbfuzzit	MicHanThi	<a href="http://mbfuzzit.sourceforge.net/">http://mbfuzzit.sourceforge.net/</a>
MySQL Connector/J	MicHanThi	<a href="http://www.mysql.com/">http://www.mysql.com/</a>
Typo3	SILVA	<a href="http://typo3.org">http://typo3.org</a>
xerces-java	MicHanThi	<a href="http://xml.apache.org/">http://xml.apache.org/</a>
<b>Bioinformatics Tools:</b>		
ARB PT Server	SILVA	<a href="http://www.arb-home/">http://www.arb-home/</a>
GenDB	MicHanThi	<a href="http://www.cebitec.uni-bielefeld.de/groups/brf/software/gendb_info/">http://www.cebitec.uni-bielefeld.de/groups/brf/software/gendb_info/</a>
InterProScan	MicHanThi	<a href="http://www.ebi.ac.uk/Tools/InterProScan/">http://www.ebi.ac.uk/Tools/InterProScan/</a>
NCBI BLAST	MicHanThi	<a href="http://blast.ncbi.nlm.nih.gov/Blast.cgi">http://blast.ncbi.nlm.nih.gov/Blast.cgi</a>
Pintail	SILVA	<a href="http://www.bioinformatics-toolkit.org/Pintail/">http://www.bioinformatics-toolkit.org/Pintail/</a>
RNAmmer	SILVA	<a href="http://www.cbs.dtu.dk/services/RNAmmer/">http://www.cbs.dtu.dk/services/RNAmmer/</a>
SignalP	MicHanThi	<a href="http://www.cbs.dtu.dk/services/SignalP/">http://www.cbs.dtu.dk/services/SignalP/</a>
TMHMM	MicHanThi	<a href="http://www.cbs.dtu.dk/services/TMHMM/">http://www.cbs.dtu.dk/services/TMHMM/</a>
<b>Databases &amp; Third Party Resources:</b>		
DSMZ nomenclature	SILVA	<a href="http://www.dsmz.de/microorganisms/main.php?contentleft_id=14">http://www.dsmz.de/microorganisms/main.php?contentleft_id=14</a>
EMBL	SILVA	<a href="http://www.ebi.ac.uk/embl/">http://www.ebi.ac.uk/embl/</a>

Continued on next page

Table A.1 – continued from previous page

<b>Name</b>	<b>Used In</b>	<b>URL</b>
EMBL EMVEC	SILVA	<a href="http://www.ebi.ac.uk/Tools/blastall/vectors.html">http://www.ebi.ac.uk/Tools/blastall/vectors.html</a>
EnvDB	SILVA	<a href="http://metagenomics.uv.es/envDB/">http://metagenomics.uv.es/envDB/</a>
Greengenes	SILVA	<a href="http://greengenes.lbl.gov/">http://greengenes.lbl.gov/</a>
InterPro	MicHanThi	<a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a>
Livingtree	SILVA	<a href="http://www.arb-silva.de/projects/living-tree/">http://www.arb-silva.de/projects/living-tree/</a>
megx	SILVA	<a href="http://www.megx.net">http://www.megx.net</a>
NCBI nr	MicHanThi	<a href="ftp://ftp.ncbi.nih.gov/blast/db/">ftp://ftp.ncbi.nih.gov/blast/db/</a>
NCBI UniVec	SILVA	<a href="http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html">http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html</a>
RDP	SILVA	<a href="http://rdp.cme.msu.edu/">http://rdp.cme.msu.edu/</a>
StrainInfo	SILVA	<a href="http://www.straininfo.net/">http://www.straininfo.net/</a>
SwissProt	MicHanThi	<a href="http://www.expasy.ch/sprot/">http://www.expasy.ch/sprot/</a>

## Appendix B

# MicHanThi Rule Base & SILVA Meta Data

**Table B.1:** *Rule base used to evaluate the reliability of BLAST observations (10). The rules used to rate InterProScan observations map the four possible values for the linguistic variable E-value to the four reliability classes.*

	E-value		Query / Target Coverage				Conclusion
<b>Rule1</b>	unreliable	&	none	&	none	→	bad
<b>Rule2</b>	unreliable	&	none	&	partial	→	bad
<b>Rule3</b>	unreliable	&	none	&	complete	→	bad
<b>Rule4</b>	unreliable	&	partial	&	none	→	bad
<b>Rule5</b>	unreliable	&	partial	&	partial	→	bad
<b>Rule6</b>	unreliable	&	partial	&	complete	→	bad
<b>Rule7</b>	unreliable	&	complete	&	none	→	bad
<b>Rule8</b>	unreliable	&	complete	&	partial	→	bad
<b>Rule9</b>	unreliable	&	complete	&	complete	→	bad
<b>Rule10</b>	uncertain	&	none	&	none	→	bad
<b>Rule11</b>	uncertain	&	none	&	partial	→	bad
<b>Rule12</b>	uncertain	&	none	&	complete	→	bad
<b>Rule13</b>	uncertain	&	partial	&	none	→	bad
<b>Rule14</b>	uncertain	&	partial	&	partial	→	average
<b>Rule15</b>	uncertain	&	partial	&	complete	→	average
<b>Rule16</b>	uncertain	&	complete	&	none	→	bad
<b>Rule17</b>	uncertain	&	complete	&	partial	→	average
<b>Rule18</b>	uncertain	&	complete	&	complete	→	average
<b>Rule19</b>	reliable	&	none	&	none	→	average
<b>Rule20</b>	reliable	&	none	&	partial	→	average
<b>Rule21</b>	reliable	&	none	&	complete	→	average
<b>Rule22</b>	reliable	&	partial	&	none	→	average
<b>Rule23</b>	reliable	&	partial	&	partial	→	average
<b>Rule24</b>	reliable	&	partial	&	complete	→	good
<b>Rule25</b>	reliable	&	complete	&	none	→	average
<b>Rule26</b>	reliable	&	complete	&	partial	→	average
<b>Rule27</b>	reliable	&	complete	&	complete	→	good
<b>Rule28</b>	very_reliable	&	none	&	none	→	average
<b>Rule29</b>	very_reliable	&	none	&	partial	→	average
<b>Rule30</b>	very_reliable	&	none	&	complete	→	average
<b>Rule31</b>	very_reliable	&	partial	&	none	→	average
<b>Rule32</b>	very_reliable	&	partial	&	partial	→	good
<b>Rule33</b>	very_reliable	&	partial	&	complete	→	very_good
<b>Rule34</b>	very_reliable	&	complete	&	none	→	average
<b>Rule35</b>	very_reliable	&	complete	&	partial	→	good
<b>Rule36</b>	very_reliable	&	complete	&	complete	→	very_good

**Table B.2:** Meta data exported into the ARB database files and their sources. EMBL field names that start with the '/' character are parsed from the EMBL feature table and field names that have two capital letters are parsed from the EMBL header. Sources denoted as SILVA are either values calculated by the SILVA binaries or imported from third party sources into the SILVA databases. env entries originate from the EnvDB (<http://metagenomics.wi.es/envDB/>) database. This data was kindly provided by Renzo Kottman (Microbial Genomics Group – Mack Planck Institute for Marine Microbiology).

ARB Field Name	Source	Source Field Name	Description
ARB_color	ARB		Stores the information about sequence colours
DOC_slv	env	doc	Dissolved organic carbon concentration in the environment at time of sampling
POC_slv	env	poc	Particulate Organic Carbon concentration in the environment at time of sampling
acc	EMBL	ID	Accession Number
ali_xx/data	EMBL	SQ	(Aligned) sequence data
align_bp_score_slv	SILVA		Calculates the number of bases in helices in the aligned sequence taken into account canonical and non canonical basepairing. The cost matrix is taken from ARB Probe_Match 2
align_cutoff_head_slv	SILVA		Unaligned bases at the beginning of the sequence
align_cutoff_tail_slv	SILVA		Unaligned bases at the end of the sequence
align_log_slv	SILVA		Indicates if the sequence was reversed and/or complemented
align_quality_slv	SILVA		Maximal similarity to reference sequence in the seed
aligned	user		User defined entry, e.g. name and date of the person who aligned the sequence
aligned_slv	SILVA		Data and time of alignment by Silva
alternative_name_slv	SILVA	synonym	Synonyms or basonyms of the species according to the DSMZ 'nomenclature up to date' catalogue
altitude_slv	env	altitude	The altitude of sampling location above sea level
ambig	ARB		Ambiguities calculated in ARB using count ambiguities
ambig_slv	SILVA		Calculated percent ambiguities in the sequences, a maximum of 2% is allowed
ann_src_slv	SILVA	field	Additional sources of sequence information is indicated in this field. Current identifiers: RNAmmer and RDP
author	EMBL	RA	Reference authors
bio_material	EMBL	/bio_material	Identifier for the biological material from which the nucleic acid sequenced was obtained
chlorophyll_slv	env	chlorophyll	Chlorophyll concentration in the environment at time of sampling
clone	EMBL	/clone	Cone from which the sequence was obtained
clone_lib	EMBL	/clone_lib	Clone library from which the sequence was obtained
collected_by	EMBL	/collected_by	Name of the person who collected the specimen
collection_date	EMBL	/collection_date	Date that the sample/specimen was collected
collection_time_slv	env	collection_time	Time that the sample was collected in hours and minutes (formerly sampling_time_slv)
country	EMBL	/country	Geographical origin of sequenced sample
culture_collection	EMBL	/culture_collection	Institution code and identifier for the culture from which the nucleic acid sequenced was obtained, with optional collection code

Continued on next page



Table B.2 – continued from previous page

ARB Field Name	Source	Source Field Name	Description
date	EMBL	D/T	Entry creation and update date separated by
description	EMBL	DE	Description
dissolved_oxygen_slv	env	dissolved_oxygen	Dissolved oxygen concentration in the environment at time of sampling
env_sample	EMBL	/environmental_sample	Identifies sequences derived by direct molecular isolation from a bulk environmental DNA sample (by PCR with or without subsequent cloning of the product, DGGE, or other anonymous methods) with no reliable identification of the source organism
full_name	EMBL	OS	Organism species
gene	EMBL	{gene,note,product}	Symbol of the gene corresponding to a sequence region
geodetic_datum_slv	env	geodetic_datum	Geodetic datum e.g. WGS 84
insdc	EMBL	PR	The International Nucleotide Sequence Database Collaboration (INSDC) Project Identifier that has been assigned to the entry
habitat_slv	env	habitat	Description of the habitat, like marine, freshwater etc.
homop_events_slv	SILVA		Absolute number of repetitive elements with more than four bases
homop_slv	SILVA		Calculated percentages repetitive bases with more than four bases, a maximum of 2% is allowed
isolate	EMBL	/isolate	Individual isolate from which the sequence was obtained
isolation_source	EMBL	/isolation_source	Describes the physical, environmental and/or local geographical source of the biological sample from which the sequence was derived
journal	EMBL	RL	Geographical coordinates of the location where the specimen was collected
lat_lon	EMBL	/lat_lon	Details of the measurement of geographic coordinates, like: Was latitude and longitude measured by GPS, derived from map, retrieved from literature?
lat_lon_details_slv	env	lat_lon_details	Identifies sequences from a culture-independent genomic analysis of an environmental sample submitted as part of a whole genome shotgun project. Contains original predictions (EMBL) and RNAmmer calls.
metagenomic_slv	SILVA		
mol_type	EMBL	/mol_type	Internal ARB database ID, do not change
name	ARB		Nitrate concentration in the environment at time of sampling
nitrate_slv	env	nitrate	Number of nucleotides calculated by ARB using 'count nucleotides'
nuc	ARB		Aligned bases within gene boundaries
nuc_gene_slv	SILVA		Identifies the biological source of the specified span of the sequence
nuc_region	EMBL	FT start..stop	Reference positions
nuc_rp	EMBL	RP	Number of nucleotides coding for the respective rRNA gene
nuc_term	ARB		pH value in the environment at time of sampling
pH_slv	env	pH	PCR primers that were used to amplify the sequence.
pcr-primers	EMBL	/pcr-primers	
phosphate_slv	env	phosphate	
pintail_slv	SILVA		Information about potential sequence anomalies detected by Pintail (1)
product	EMBL	/product	Name of the product associated with the feature

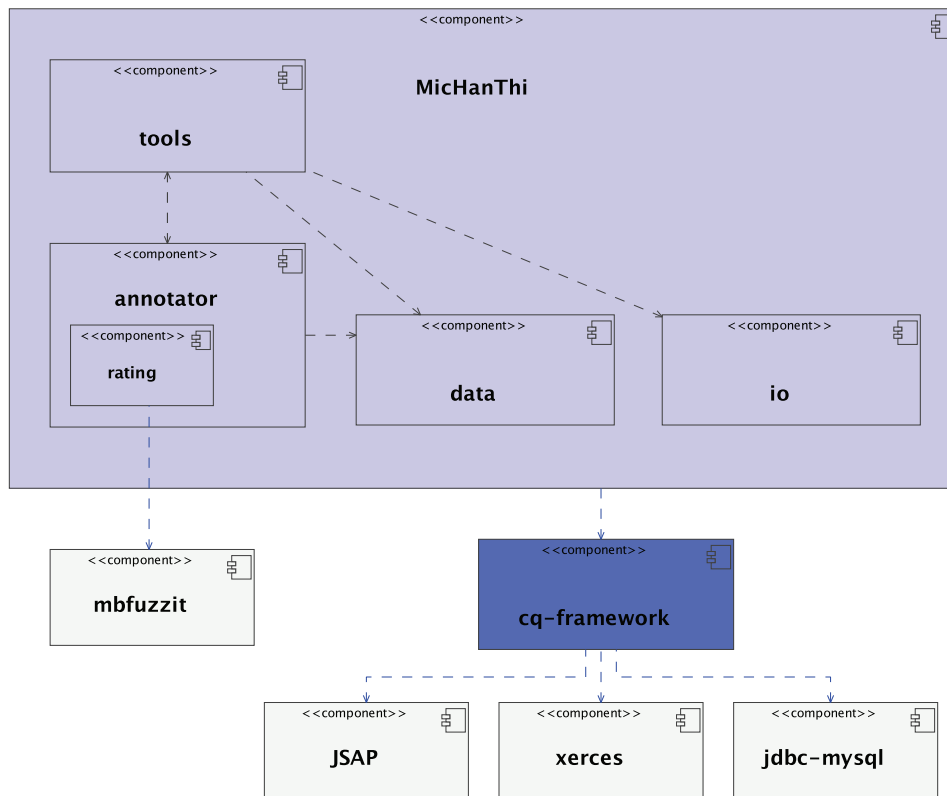
Continued on next page

Table B.2 – continued from previous page

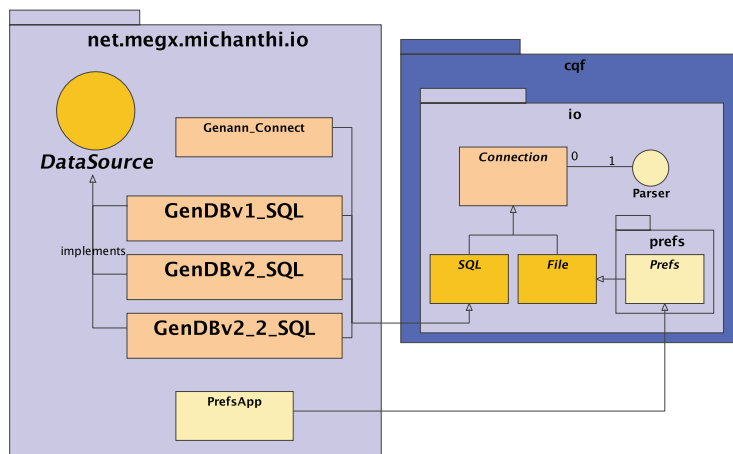
ARB Field Name	Source	Source Field Name	Description
project_name_slv	SILVA	project_name	Name of the sequencing project
publication_doi	EMBL	RX	Cross-reference DOI number
pubmed_id	EMBL	RX	Cross-reference Pubmed ID
remark	user		Free for remarks
salinity_slv	env	salinity	Salinity concentration in the environment at time of sampling
sample_identifier_slv	env	sample_identifier	A unique identifier (ID) given to the sample that allows to cross-reference samples and contextual data
sample_material_slv	env	sample_material	Describes the sample material that was collected, e.g. water, sediment, biofilm, vent fluid etc.
sample_volume_slv	env	sample_size	Volume of the sample that was collected
sediment_depth_slv	env	sediment_depth	Depth of the sediment from where the sample was collected
seq_quality_slv	SILVA		Summary sequence quality value calculated based on values from vector, ambiguities and homopolymers, 100 means very good
silicate_slv	env	silicate	Silicate concentration in the environment at time of sampling
specific_host	EMBL	/specific_host	Natural host from which the sequence was obtained
specimen_voucher	EMBL	/specimen_voucher	An identifier of the individual or collection of the source organism and the place where it is currently stored, usually an institution
start	EMBL	FT start	Start of the ribosomal RNA gene
stop	EMBL	FT stop	Stop of the ribosomal RNA gene
strain	EMBL	/strain	Strain from which the sequence was obtained.
submit_author	EMBL	RA	Submission authors from reference location
submit_date	EMBL	RL	Submission date from reference location
tax_emb1	EMBL	OC	Organism classification according to EMBL
tax_emb1_name	EMBL	OS	Organism name taken from the classification field
tax_egg	SILVA		Taxonomy mapped from Greengenes
tax_egg_name	SILVA		Organism name in Greengenes
tax_rdp	SILVA		Nomenclatural taxonomy mapped from RDP
tax_rdp_name	SILVA		Organism name in RDP
tax_xref_emb1	EMBL	/dbxref (taxoni:)	Database cross-reference: pointer to related information in another database
temperature_slv	env	temperature	Temperature in the environment at time of sampling
title	EMBL	RT	Reference title
tmp	ARB		Used by diverse ARB modules
vector_slv	SILVA		Percent vector contamination, a maximum of 5% is allowed
version	EMBL	ID (SV)	Subversion from identification line
water_depth_slv	env	water_depth	Depth of the water column from where the sample was collected

## Appendix C

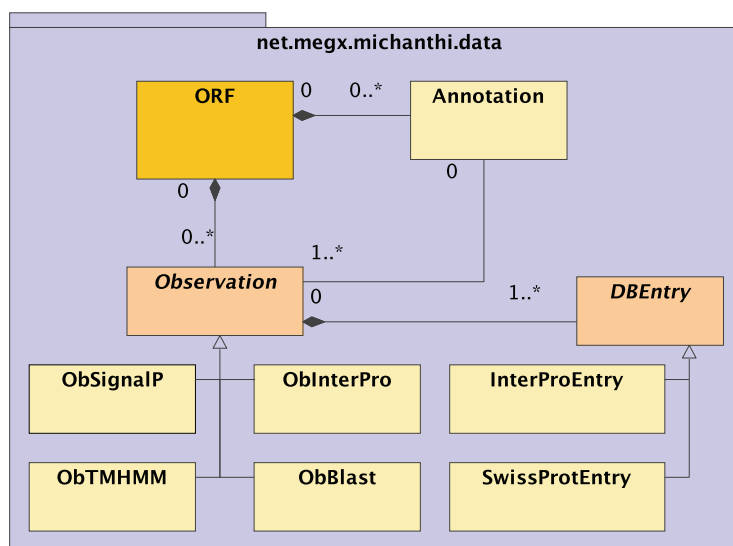
# MicHanThi Design



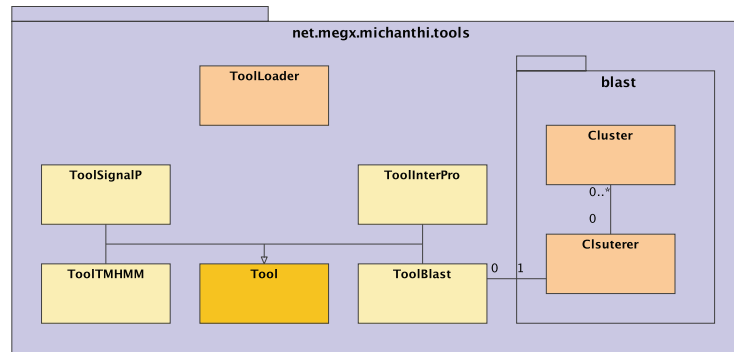
**Figure C.1:** *MicHanThi* modules overview (10). *MicHanThi* uses an abstract description of the sources of information, such as the annotation system, as well as the analysis tools. It consists of four modules: (i) the IO module (ii) the DATA module, (iii) the TOOLS module, and (iv) the ANNOTATOR module.



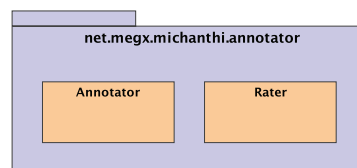
**Figure C.2:** Module IO overview (10). The IO module offers functions to access data from different sources transparently from their means of storage.



**Figure C.3:** Module DATA overview (10). The DATA module represents the information necessary to annotate an ORF. It represents the ORF, information about the ORF (observations), additional information about the observations found in the Swiss-Prot or InterPro databases, and it represents the annotations of an ORF.



**Figure C.4:** Module *TOOL* overview (10). The *TOOLS* module introduces an abstraction layer to the semantics of the different tools and it provides an interface which can be used to evaluate observations, and to create annotations.



**Figure C.5:** Module *ANNOTATOR* overview (10). The annotation process is managed by the *ANNOTATOR* module. Once the main program initialises the software and fetches all relevant information from the data source it calls this module to annotate the ORF.



# Bibliography

- [1] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, A. Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K.-H. Schleifer, "ARB: a software environment for sequence data," *Nucleic Acids Res.*, vol. 32, no. 4, pp. 1363–1371, 2004.
- [2] J. Wuyts, G. Perriere, and Y. Van de Peer, "The European ribosomal RNA database," *Nucleic Acids Res.*, vol. 32, no. suppl\_1, pp. D101–103, 2004.
- [3] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB," *Appl. Environ. Microbiol.*, vol. 72, no. 7, pp. 5069–5072, 2006.
- [4] J. R. Cole, B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje, "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis," *Nucleic Acids Res.*, vol. 33, pp. D294–D296, 2005.
- [5] R. Aziz, D. Bartels, A. Best, M. DeJongh, T. Disz, R. Edwards, K. Formsma, S. Gerdes, E. Glass, M. Kubal, F. Meyer, G. Olsen, R. Olson, A. Osterman, R. Overbeek, L. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko, "The rast server: Rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, no. 1, p. 75, 2008.
- [6] V. M. Markowitz, F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N. Ivanova, and N. C. Kyrpides, "The integrated microbial genomes (IMG) system," *Nucleic Acids Res.*, vol. 34, pp. D344–348, 2006.
- [7] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: A new generation of protein database search programs.," *FASEB J.*, vol. 12, pp. A1326–A1326, 1998.
- [8] E. M. Zdobnov and R. Apweiler, "InterProScan - an integration platform for the signature-recognition methods in InterPro," *Bioinformatics*, vol. 17, no. 9, pp. 847–848, 2001.
- [9] F. Meyer, A. Goesmann, A. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich, and A. PÄEhler, "GenDB-an open source genome annotation system for prokaryote genomes," *Nucleic Acids Res.*, vol. 31, no. 8, pp. 2187–2195, 2003.
- [10] C. Quast, "MicHanThi - Design and Implementation of a System for the Prediction of Gene Functions in Genome Annotation Projects," Master's thesis, University of Bremen, 2006.
- [11] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. A. Sigrist, and E. M. Zdobnov, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 37–40, 2001.
- [12] J. D. Watson and F. H. Crick, "A structure for Deoxyribose Nucleic Acid," *Nature*, vol. 171, pp. 737–738, April 1953.
- [13] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 74, pp. 5463–54637, December 1977.

- [14] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, and J. Merrick, "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.," *Science*, vol. 269, pp. 496–512, 1995.
- [15] R. L. Sinsheimer, "The human genome initiative," *FASEB J.*, vol. 5, no. 14, pp. 2885–, 1991.
- [16] V. Torsvik, J. Goksoyr, and F. L. Daae, "High diversity in DNA of soil bacteria.," *Appl. Environ. Microbiol.*, vol. 56, no. 3, pp. 782–787, 1990.
- [17] J. Handelsman, "Metagenomics: Application of Genomics to Uncultured Microorganisms," *Microbiol. Mol. Biol. Rev.*, vol. 68, no. 4, pp. 669–685, 2004.
- [18] J. Stein, T. Marsh, K. Wu, H. Shizuya, and E. DeLong, "Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon," *J. Bacteriol.*, vol. 178, no. 3, pp. 591–599, 1996.
- [19] R. I. Amann, W. Ludwig, and K. H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation," *Microbiol. Rev.*, vol. 59, no. 1, pp. 143–169, 1995.
- [20] E. Prüße, "Incremental approach to multiple sequence alignment using directed acyclical graphs.," Master's thesis, University of Bremen, 2007.
- [21] F. O. Glöckner, M. Kube, M. Bauer, H. Teeling, T. Lombardot, W. Ludwig, D. Gade, A. Beck, K. Borzym, K. Heitmann, R. Rabus, H. Schlesner, R. Amann, and R. Reinhardt, "Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 14, pp. 8298–8303, 2003.
- [22] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith, "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.
- [23] M. Ronaghi, "Pyrosequencing Sheds Light on DNA Sequencing," *Genome Res.*, vol. 11, pp. 3–11, 2001.
- [24] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, pp. 376–380, Sept. 2005.
- [25] D. R. Bentley, "Whole-genome re-sequencing," *Curr. Opin. Genet. Dev.*, vol. 16, no. 6, pp. 545 – 552, 2006.
- [26] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson, "A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning," *Genome Res.*, vol. 18, no. 7, pp. 1051–1063, 2008.
- [27] T. Woyke, H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Glöckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E. M. Rubin, and N. Dubilier, "Symbiosis insights through metagenomic analysis of a microbial consortium," *Nature*, vol. 443, pp. 950–955, Oct. 2006.
- [28] D. W. Mount, *Bioinformatics Sequence and Genome Analysis*. CSHL Press, second ed., 2004.
- [29] M. L. Green and P. D. Karp, "Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers," *Nucleic Acids Res.*, vol. 33, no. 13, pp. 4035–4039, 2005.
- [30] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, pp. 25–29, May 2000.
- [31] S. F. Altschul, W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman, "Basic Local Alignment Search Tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, Oct. 1990.



- [32] S. R. Eddy, "Profile hidden Markov models.," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.
- [33] L. B. Koski, M. W. Gray, B. F. Lang, and G. Burger, "AutoFACT: An Automatic Functional Annotation and Classification Tool," *BMC Bioinformatics*, vol. 6, pp. 1–11, 2005.
- [34] G. H. van Domselaar, P. Stothard, S. Shrivastava, J. A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner, and D. S. Wishart, "BASys: a web server for automated bacterial genome annotation.," *Nucleic Acids Res.*, vol. 33, pp. 455–459, 2005.
- [35] R. Overbeek, N. Larsen, T. Walunas, M. D'Souza, G. Pusch, J. Eugene Selkov, K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova, and N. Kyrpides, "The ERGO genome analysis and discovery system," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 164–171, 2003.
- [36] R. Overbeek, T. Disz, and R. Stevens, "The SEED: a peer-to-peer environment for genome annotation," *Communications of the ACM*, vol. 47, no. 11, pp. 46–51, 2004.
- [37] R. Bellman, *Dynamic Programming*. Princeton Univ Pr, June 1957.
- [38] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, March 1970.
- [39] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," *J. Mol. Biol.*, vol. 147, pp. 195–197, 1981.
- [40] D. G. Higgins and S. P. M., "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.," *Gene*, vol. 73, pp. 237–244, December 1988.
- [41] K. Katoh and H. Toh, "Recent developments in the MAFFT multiple sequence alignment program," *Brief Bioinform*, vol. 9, no. 4, pp. 286–298, 2008.
- [42] R. C. Edgar, "Muscle: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, no. 1, p. 113, 2004.
- [43] S. Karlin and S. Altschul, "Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 87, no. 6, pp. 2264–2268, 1990.
- [44] K. Okubo, H. Sugawara, T. Gojobori, and Y. Tateno, "DDBJ in preparation for overview of research activities behind data submissions," *Nucleic Acids Res.*, vol. 34, pp. D6–9, 2006.
- [45] G. H. Hamm and G. N. Cameron, "The EMBL data library," *Nucleic Acids Res.*, vol. 14, pp. 5–9, January 1986.
- [46] C. Burks, J. W. Fickett, W. B. Goad, M. Kanehisa, F. I. Lewitter, R. W. P., C. D. Swindell, T. C. S., and B. H. S., "The GenBank nucleic acid sequence database.," *Comp Appl Biosci*, vol. 1, pp. 225–233, December 1985.
- [47] A. Bairoch and B. Boeckmann, "The SWISS-PROT protein sequence data bank," *Nucleic Acids Res.*, vol. 19, pp. 2247–2249, 1991.
- [48] W. C. Barker, L. T. Hunt, D. G. George, L. S. Yeh, H. R. Chen, M. C. Blomquist, E. I. Seibel-Ross, A. Elzanowski, B. J. K., and F. D. A. et al., "Protein sequence database of the protein identification resource (PIR).," *Protein Seq. Data Anal.*, vol. 1, no. 1, pp. 43–49, 1987.
- [49] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.
- [50] K. D. Pruitt, K. S. Katz, H. Sicotte, and D. R. Maglott, "Introducing RefSeq and LocusLink: curated human genome resources at the NCBI," *Trends Genet.*, vol. 16, pp. 44–47, January 2000.
- [51] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res.*, vol. 32, no. 9, pp. D115–119, 2004.
- [52] R. D. Finn, J. Tate, J. Mistry, P. C. Coghill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Res.*, vol. 36, no. suppl.1, pp. D281–288, 2008.

- [53] D. H. Haft, B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White, "TIGRFAMs: a protein family resource for the functional identification of proteins," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 41–43, 2001.
- [54] F. Corpet, J. Gouzy, and D. Kahn, "The ProDom database of protein domain families," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 323–326, 1998.
- [55] B. L. Maidak, J. R. Cole, T. G. Lilburn, J. Parker, Charles T., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje, "The RDP-II (Ribosomal Database Project)," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 173–174, 2001.
- [56] J. R. Cole, B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje, "The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data," *Nucleic Acids Res.*, vol. 35, no. suppl.1, pp. D169–172, 2007.
- [57] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy, "Infernal 1.0: inference of RNA alignments," *Bioinformatics*, vol. 25, no. 10, pp. 1335–1337, 2009.
- [58] J. DeSantis, T. Z., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen, "NASt: a multiple sequence alignment server for comparative analysis of 16S rRNA genes," *Nucleic Acids Res.*, vol. 34, no. suppl.2, pp. W394–399, 2006.
- [59] K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman, "At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies," *Appl. Environ. Microbiol.*, vol. 71, no. 12, pp. 7724–7736, 2005.
- [60] T. Huber, G. Faulkner, and P. Hugenholtz, "Bellerophon: a program to detect chimeric sequences in multiple sequence alignments," *Bioinformatics*, vol. 20, no. 14, pp. 2317–2319, 2004.
- [61] M. Bauer, M. Kube, H. Teeling, M. Richter, T. Lombardot, E. Allers, C. A. Würdemann, C. Quast, H. Kuhl, F. Knaust, D. Woebken, K. Bischof, M. Mussmann, J. V. Choudhuri, F. Meyer, R. Reinhardt, R. I. Amann, and F. O. Glöckner, "Whole genome analysis of the marine *Bacteroidetes Gramella forsetii* reveals adaptations to degradation of polymeric organic matter," *Environ. Microbiol.*, vol. 8, pp. 2201–2213, October 2006.
- [62] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne., "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *Protein Eng.*, vol. 10, no. 1, pp. 1–6, 1997.
- [63] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes," *J. Mol. Biol.*, vol. 305, no. 3, pp. 567 – 580, 2001.
- [64] T. Lombardot, R. Kottmann, G. Giuliani, A. de Bono, N. Addor, and F. Glöckner, "Metalook: a 3d visualisation software for marine ecological genomics," *BMC Bioinformatics*, vol. 8, no. 1, p. 406, 2007.
- [65] U. Bohnebeck, T. Lombardot, R. Kottmann, and F. O. Glöckner, "Metamine - a tool to detect and analyse gene patterns in their environmental context," *BMC Bioinformatics*, vol. 9, no. 1, p. 459, 2008.
- [66] T. Lombardot, R. Kottmann, H. Pfeffer, M. Richter, H. Teeling, C. Quast, and F. O. Glöckner, "Megx.net—database resources for marine ecological genomics," *Nucleic Acids Res.*, vol. 34, pp. D390–393, 2006.
- [67] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. M. Zhu, and R. Apweiler, "The embl nucleotide sequence database," *Nucleic Acids Res.*, vol. 33, pp. D29–D33, 2005.
- [68] P. Yarza, M. Richter, J. Peplies, J. Euzéby, R. Amann, K.-H. Schleifer, W. Ludwig, F. O. Glöckner, and R. Rosselló-Móra, "The all-species living tree project: A 16s rRNA-based phylogenetic tree of all sequenced type strains," *System. Appl. Microbiol.*, vol. 31, no. 4, pp. 241 – 250, 2008.
- [69] K. Lagesen, P. Hallin, E. A. Rodland, H.-H. Staerfeldt, T. Rognes, and D. W. Ussery, "RNAmmer: consistent and rapid annotation of ribosomal RNA genes," *Nucleic Acids Res.*, vol. 35, no. 9, pp. 3100–3108, 2007.

- [70] P. Romano, P. Dawyndt, F. Piersigilli, and J. Swings, "Improving interoperability between microbial information and sequence databases," *BMC Bioinformatics*, vol. 6, no. Suppl 4, p. S23, 2005.
- [71] J. Peplies, R. Kottmann, W. Ludwig, and F. O. Glöckner, "A standard operating procedure for phylogenetic inference (soppi) using (rrna) marker genes," *System. Appl. Microbiol.*, vol. 31, no. 4, pp. 251–257, 2008.
- [72] J. Celko, *Trees and Hierarchies in SQL for Smarties*. Morgan Kaufmann, 3 ed., 2004.
- [73] B. M. Fuchs, S. Spring, H. Teeling, C. Quast, J. Wulf, M. Schattner, S. Yan, S. Ferriera, J. Johnson, F. O. Glöckner, and R. Amann, "Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, no. 8, pp. 2891–2896, 2007.
- [74] M. Richter, M. Kube, D. A. Bazylinski, T. Lombardot, F. O. Glöckner, R. Reinhardt, and D. Schuler, "Comparative Genome Analysis of Four Magnetotactic Bacteria Reveals a Complex Set of Group-Specific Genes Implicated in Magnetosome Biomineralization and Function," *J. Bacteriol.*, vol. 189, no. 13, pp. 4899–4910, 2007.
- [75] D. Woeckel, H. Teeling, P. Wecker, A. Dumitriu, I. Kostadinov, E. F. DeLong, R. Amann, and F. O. Glöckner, "Fosmids of novel marine planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes," *ISME J*, vol. 1, pp. 419–435, August 2007.
- [76] M. Mußmann, F. Z. Hu, M. Richter, D. de Beer, A. Preisler, B. B. Jørgensen, M. Huntemann, F. O. Glöckner, R. Amann, W. J. H. Koopman, R. S. Lasken, B. Janto, J. Hogg, P. Stoodley, R. Boissy, and G. D. Ehrlich, "Insights into the genome of large sulfur bacteria revealed by analysis of single filaments," *PLoS Biol.*, vol. 5, p. e230, 08 2007.
- [77] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "Real-Time DNA Sequencing from Single Polymerase Molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.
- [78] E. R. Mardis, "Anticipating the \$1,000 genome," *Genome Biol.*, vol. 7, no. 7, p. 112, 2006.
- [79] R. F. Service, "GENE SEQUENCING: The Race for the \$1000 Genome," *Science*, vol. 311, no. 5767, pp. 1544–1546, 2006.
- [80] L. Kedes, "Genomics prize—the X PRIZE Foundation. Interview by Vicki Glaser.," *Rejuvenation Res.*, vol. 10, pp. 237–42, June 2007.
- [81] A. E. Darling, L. Carey, and W. Chun Feng, "The Design, Implementation, and Evaluation of mpiBLAST," in *4th International Conference on Linux Clusters: The HPC Revolution 2003 in conjunction with ClusterWorld Conference & Expo*, June 2003.
- [82] Y. Sun, Y. Cai, L. Liu, F. Yu, M. L. Farrell, W. McKendree, and W. Farmerie, "ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences," *Nucleic Acids Res.*, vol. 37, no. 10, pp. e76–, 2009.
- [83] J. Wilkening, N. Desai, F. Meyer, and A. Wilke, "Using clouds for metagenomics – A case study," in *2009 IEEE International Conference on Cluster Computing (Cluster 2009)*, 2009.
- [84] G. E. FOX, K. R. PECHMAN, and C. R. WOESE, "Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Prokaryotic Systematics," *Int. J. Syst. Bacteriol.*, vol. 27, no. 1, pp. 44–57, 1977.
- [85] N. R. Pace, D. A. Stahl, G. J. Olsen, and D. J. Lane, "Analyzing natural microbial populations by rRNA sequences," *ASM News*, vol. 51, pp. 4–12, 1985.
- [86] G. J. Olsen, D. J. Lane, S. J. Giovannoni, N. R. Pace, and D. A. Stahl, "Microbial ecology and evolution: a ribosomal rRNA approach," *Annu. Rev. Microbiol.*, vol. 40, pp. 337–65, 1986.
- [87] S. J. Giovannoni, E. F. DeLong, G. J. Olsen, and N. R. Pace, "Phylogenetic group-specific oligodeoxynucleotide probes for identification of single microbial cells," *J. Bacteriol.*, vol. 170, pp. 720–726, 1988.

- [88] D. M. Ward, R. Weller, and M. M. Bateson, "16s rna sequences reveal numerous uncultured microorganisms in a natural community," *Nature*, vol. 345, no. 6270, pp. 63–65, 1990.
- [89] N. Pace, "A molecular view of microbial diversity and the biosphere," *Science*, vol. 276, pp. 734–740, 1997.
- [90] W. Ludwig and K. H. Schleifer, "Molecular phylogeny of bacteria based on comparative sequence analysis of conserved genes," in *Microbial phylogeny and evolution, concepts and controversies* (J. Sapp, ed.), pp. 70–98, New York: Oxford university press, 2005.
- [91] J. Peplies, F. O. Glöckner, R. Amann, and W. Ludwig, "Comparative sequence analysis and oligonucleotide probe design based on 23s rna genes of alphaproteobacteria from north sea bacterioplankton," *System. Appl. Microbiol.*, vol. 27, no. 5, pp. 573–580, 2004.
- [92] J. Wuyts, P. De Rijk, Y. Van de Peer, T. Winkelmans, and R. De Wachter, "The european large subunit ribosomal rna database," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 175–177, 2001.
- [93] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. T. Chen, S. B. Dewell, A. de Winter, J. Drake, L. Du, J. M. Fierro, R. Forte, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, S. K. Hutchison, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, W. L. Lee, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. Reifler, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, D. A. Willoughby, P. G. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors (vol 437, pg 376, 2005)," *Nature*, vol. 441, no. 7089, pp. 120–120, 2006.
- [94] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, "Microbial diversity in the deep sea and the underexplored "rare biosphere"," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 32, pp. 12115–12120, 2006.
- [95] C. Lee, C. Grasso, and M. F. Sharlow, "Multiple sequence alignment using partial order graphs," *Bioinformatics*, vol. 18, no. 3, pp. 452–464, 2002.
- [96] T. Z. DeSantis, I. Dubosarskiy, S. R. Murray, and G. L. Andersen, "Comprehensive aligned sequence construction for automated design of effective probes (cascade-p) using 16s rdna," *Bioinformatics*, vol. 19, no. 12, pp. 1461–1468, 2003.
- [97] R. R. Gutell, N. Larsen, and C. R. Woese, "Lessons from an evolving rna: 16s and 23s rna structures from a comparative perspective," *Microbiol. Rev.*, vol. 58, no. 1, pp. 10–26, 1994.
- [98] J. R. Marchesi, T. Sato, A. J. Weightman, T. A. Martin, J. C. Fry, S. J. Hiom, and W. G. Wade, "Design and evaluation of useful bacterium-specific pcr primers that amplify genes coding for bacterial 16s rna," *Appl. Environ. Microbiol.*, vol. 64, no. 2, pp. 795–799, 1998.
- [99] G. Muyzer, E. de Waal, and A. Uitterlinden, "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16s rna," *Appl. Environ. Microbiol.*, vol. 59, no. 3, pp. 695–700, 1993.
- [100] E. F. DeLong, "Archaea in coastal marine environments," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 89, no. 12, pp. 5685–5689, 1992.
- [101] J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, "Balibase 3.0: Latest developments of the multiple sequence alignment benchmark," *Proteins Struct. Funct. Bioinform.*, vol. 61, no. 1, pp. 127–136, 2005.
- [102] D. Field, G. Garrity, T. Gray, J. Selengut, P. Sterk, N. Thomson, T. Tatusova, G. Cochrane, F. O. Glöckner, R. Kottmann, A. L. Lister, Y. Tateno, and R. Vaughan, "egenomics: Cataloguing our complete genome collection iii," *Comp. Funct. Genomics*, vol. 2007, pp. 1–7, 2007.
- [103] S. H. Hong, J. Bunge, S. O. Jeon, and S. S. Epstein, "Predicting microbial species richness," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 1, pp. 117–122, 2006.
- [104] C. Pedros-Alio, "Marine microbial diversity: can it be determined?," *Trends Microbiol.*, vol. 14, no. 6, pp. 257–263, 2006.
- [105] D. Tautz, P. Arctander, A. Minelli, R. H. Thomas, and A. P. Vogler, "Dna points the way ahead of taxonomy - in assessing new approaches, it's time for dna's unique contribution to take a central role," *Nature*, vol. 418, no. 6897, pp. 479–479, 2002.

- [106] E. R. Mardis, "Next-generation dna sequencing methods," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, no. 1, pp. 387–402, 2008.
- [107] P. Stothard and D. S. Wishart, "Automated bacterial genome analysis and annotation," *Curr. Opin. Microbiol.*, vol. 9, no. 5, pp. 505 – 510, 2006.
- [108] L. A. Zadeh, "Fuzzy Logic," *IEEE*, vol. 88, pp. 83–93, 1988.
- [109] I. N. McCave, "Vertical flux of particles in the ocean," *Deep-Sea Research*, vol. 22, pp. 491–502, 1975.
- [110] A. Engel, S. Thoms, U. Riebesell, E. Rochelle-Newall, and I. Zondervan, "Polysaccharide aggregation as a potential sink of marine dissolved organic carbon," *Nature*, vol. 428, no. 6986, pp. 929–932, 2004.
- [111] A. L. Shanks and J. D. Trent, "Marine snow - sinking rates and potential role in vertical flux," *Deep-Sea Research Part a-Oceanographic Research Papers*, vol. 27, no. 2, pp. 137–143, 1980.
- [112] M. W. Silver and A. L. Alldredge, "Bathypelagic marine snow - deep-sea algal and detrital community," *J. Mar. Res.*, vol. 39, no. 3, pp. 501–530, 1981.
- [113] K. E. Kohfeld, C. L. Quere, S. P. Harrison, and R. F. Anderson, "Role of marine biology in glacial-interglacial co2 cycles," *Science*, vol. 308, no. 5718, pp. 74–78, 2005.
- [114] D. C. Smith, M. Simon, A. L. Alldredge, and F. Azam, "Intense hydrolytic enzyme-activity on marine aggregates and implications for rapid particle dissolution," *Nature*, vol. 359, no. 6391, pp. 139–142, 1992.
- [115] E. F. DeLong, D. G. Franks, and A. L. Alldredge, "Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages," *Limnol. Oceanogr.*, vol. 38, no. 5, pp. 924–934, 1993.
- [116] J. Rath, K. Y. Wu, G. J. Herndl, and E. F. DeLong, "High phylogenetic diversity in a marine-snow-associated bacterial assemblage," *Aquat. Microb. Ecol.*, vol. 14, no. 3, pp. 261–269, 1998.
- [117] L. B. Fandino, L. Riemann, G. F. Steward, R. A. Long, and F. Azam, "Variations in bacterial community structure during a dinoflagellate bloom analyzed by dgge and 16s rdna sequencing," *Aquat. Microb. Ecol.*, vol. 23, no. 2, pp. 119–130, 2001.
- [118] L. B. Fandino, L. Riemann, G. F. Steward, and F. Azam, "Population dynamics of cytophaga-flavobacteria during marine phytoplankton blooms analyzed by real-time quantitative pcr," *Aquat. Microb. Ecol.*, vol. 40, no. 3, pp. 251–257, 2005.
- [119] M. T. Cottrell and D. L. Kirchman, "Natural assemblages of marine proteobacteria and members of the cytophaga-flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter," *Appl. Environ. Microbiol.*, vol. 66, no. 4, pp. 1692–1697, 2000.
- [120] D. L. Kirchman, "The ecology of cytophaga-flavobacteria in aquatic environments," *FEMS Microbiol. Ecol.*, vol. 39, no. 2, pp. 91–100, 2002.
- [121] R. Benner, "Molecular indicators of the bioavailability of dissolved organic matter," in *Aquatic ecosystems: interactivity of dissolved organic matter* (S. Findlay and R. L. Sinsabaugh, eds.), Aquatic Ecology Series, pp. 316–342, San Diego, CA: Academic Press, 2003.
- [122] M. T. Cottrell, L. Y. Yu, and D. L. Kirchman, "Sequence and expression analysis of cytophaga-like hydrolases in a western arctic metagenomic library and the sargasso sea," *Appl. Environ. Microbiol.*, vol. 71, no. 12, pp. 8506–8513, 2005.
- [123] J. J. Grzyski, B. J. Carter, E. F. DeLong, R. A. Feldman, A. Ghadiri, and A. E. Murray, "Comparative genomics of dna fragments from six antarctic marine planktonic bacteria," *Appl. Environ. Microbiol.*, vol. 72, no. 2, pp. 1532–1541, 2006.
- [124] H. Eilers, J. Pernthaler, J. Peplies, F. O. Glöckner, G. Gerdt, and R. Amann, "Isolation of novel pelagic bacteria from the german bight and their seasonal contributions to surface picoplankton," *Appl. Environ. Microbiol.*, vol. 67, no. 11, pp. 5134–5142, 2001.
- [125] S. Hou, J. H. Saw, K. S. Lee, T. A. Freitas, C. Belisle, Y. Kawarabayasi, S. P. Donachie, A. Pikina, M. Y. Galperin, E. V. Koonin, K. S. Makarova, M. V. Omelchenko, A. Sorokin, Y. I. Wolf, Q. X. Li, Y. S. Keum, S. Campbell, J. Denery, S.-I. Aizawa, S. Shibata, A. Malahoff, and M. Alam, "Genome sequence of the deep-sea gamma-proteobacterium *Idiomarina loihiensis* reveals amino acid fermentation as a source of carbon and energy," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 52, pp. 18036–18041, 2004.

- [126] L. Riemann and F. Azam, "Widespread n-acetyl-d-glucosamine uptake among pelagic marine bacteria and its ecological implications," *Appl. Environ. Microbiol.*, vol. 68, no. 11, pp. 5554–62, 2002.
- [127] Y. Obayashi and S. Suzuki, "Proteolytic enzymes in coastal surface seawater: Significant activity of endopeptidases and exopeptidases," *Limnol. Oceanogr.*, vol. 50, no. 2, pp. 722–726, 2005.
- [128] C. Arnosti, S. Durkin, and W. H. Jeffrey, "Patterns of extracellular enzyme activities among pelagic marine microbial communities: implications for cycling of dissolved organic carbon," *Aquat. Microb. Ecol.*, vol. 38, no. 2, pp. 135–145, 2005.
- [129] J. A. Shipman, J. E. Berleman, and A. A. Salyers, "Characterization of four outer membrane proteins involved in binding starch to the cell surface of bacteroides thetaiotaomicron," *J. Bacteriol.*, vol. 182, no. 19, pp. 5365–5372, 2000.
- [130] M. A. Moran, A. Buchan, J. M. Gonzalez, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B. Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Ye, T. R. Miller, J. Carlton, D. A. Rasko, I. T. Paulsen, Q. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S. Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut, and N. Ward, "Genome sequence of silicibacter pomeroyi reveals adaptations to the marine environment," *Nature*, vol. 432, no. 7019, pp. 910–913, 2004.
- [131] S. J. Giovannoni, H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappe, J. M. Short, J. C. Carrington, and E. J. Mathur, "Genome streamlining in a cosmopolitan oceanic bacterium," *Science*, vol. 309, no. 5738, pp. 1242–1245, 2005.
- [132] N. Kennerknecht, H. Sahm, M. R. Yen, M. Patek, M. H. Saier, and L. Eggeling, "Export of l-isoleucine from corynebacterium glutamicum: A two-gene-encoded member of a new translocator family," *J. Bacteriol.*, vol. 184, no. 14, pp. 3947–3956, 2002.
- [133] B. Winnen, R. N. Hvorup, and M. H. Saier, "The tripartite tricarboxylate transporter (ttt) family," *Res. Microbiol.*, vol. 154, no. 7, pp. 457–465, 2003.
- [134] J. Xu, H. C. Chiang, M. K. Bjursell, and J. I. Gordon, "Message from a human gut symbiont: sensitivity is a prerequisite for sharing," *Trends Microbiol.*, vol. 12, no. 1, pp. 21–28, 2004.
- [135] Pinhassi, Jarone and Sala, Maria Montserrat and Havskum, Harry and Peters, Francesc and Guadayol, Òscar and Malits, Andrea and Marrase, Celia, "Changes in bacterioplankton composition under different phytoplankton regimens," *Appl. Environ. Microbiol.*, vol. 70, no. 11, pp. 6753–6766, 2004.
- [136] R. Koenig, "TonB-dependent trans-envelope signalling: the exception or the rule?," *Trends Microbiol.*, vol. 13, no. 8, pp. 343–347, 2005.
- [137] M. Simon, F. O. Glöckner, and R. Amann, "Different community structure and temperature optima of heterotrophic picoplankton in various regions of the southern ocean," *Aquat. Microb. Ecol.*, vol. 18, no. 3, pp. 275–284, 1999.
- [138] J. S. Covert and M. A. Moran, "Molecular characterization of estuarine bacterial communities that use high- and low-molecular weight fractions of dissolved organic carbon," *Aquat. Microb. Ecol.*, vol. 25, pp. 127–139, 2001.
- [139] J. A. Klappenbach, J. M. Dunbar, and T. M. Schmidt, "rrna operon copy number reflects ecological strategies of bacteria," *Appl. Environ. Microbiol.*, vol. 66, no. 4, pp. 1328–1333, 2000.
- [140] G. P. Ferguson, S. Totemeyer, M. J. MacLean, and I. R. Booth, "Methylglyoxal production in bacteria: suicide or survival?," *Arch. Microbiol.*, vol. 170, no. 4, pp. 209–219, 1998.
- [141] J. D. Tolli, S. M. Sievert, and C. D. Taylor, "Unexpected diversity of bacteria capable of carbon monoxide oxidation in a coastal marine environment, and contribution of the roseobacter-associated clade to total co oxidation," *Appl. Environ. Microbiol.*, vol. 72, no. 3, pp. 1966–1973, 2006.
- [142] C. Cosseau and J. Batut, "Genomics of the cconoqp-encoded cbb(3) oxidase complex in bacteria," *Arch. Microbiol.*, vol. 181, no. 2, pp. 89–96, 2004.
- [143] R. S. Pitcher and N. J. Watmough, "The bacterial cytochrome cbb(3) oxidases," *Biochim. Biophys. Acta, Bioenerg.*, vol. 1655, no. 1-3, pp. 388–399, 2004.

- [144] K. U. Vollack and W. G. Zumft, "Nitric oxide signaling and transcriptional control of denitrification genes in *Pseudomonas stutzeri*," *J. Bacteriol.*, vol. 183, no. 8, pp. 2516–2526, 2001.
- [145] R. K. Poole and M. N. Hughes, "New functions for the ancient globin family: bacterial responses to nitric oxide and nitrosative stress," *Mol. Microbiol.*, vol. 36, no. 4, pp. 775–783, 2000.
- [146] C. Brochier, P. Lopez-Garcia, and D. Moreira, "Horizontal gene transfer and archaeal origin of deoxyhypusine synthase homologous genes in bacteria," *Gene*, vol. 330, pp. 169–176, 2004.
- [147] R. Hosoya and K. Hamana, "Distribution of two triamines, spermidine and homospermidine, and an aromatic amine, 2-phenylethylamine, within the phylum bacteroidetes," *J. Gen. Appl. Microbiol.*, vol. 50, no. 5, pp. 255–260, 2004.
- [148] N. Nurhayati and D. Ober, "Recruitment of alkalooid-specific homospermidine synthase (hss) from ubiquitous deoxyhypusine synthase: Does *Crotalaria* possess a functional hss that still has dhs activity?," *Phytochemistry*, vol. 66, pp. 1346–1357, 2005.
- [149] S. Kopriva, T. Buchert, G. Fritz, M. Suter, R. D. Benda, V. Schunemann, A. Koprivova, P. Schurmann, A. X. Trautwein, P. M. H. Kroneck, and C. Brunold, "The presence of an iron-sulfur cluster in adenosine 5'-phosphosulfate reductase separates organisms utilizing adenosine 5'-phosphosulfate and phosphoadenosine 5'-phosphosulfate for sulfate assimilation," *J. Biol. Chem.*, vol. 277, no. 24, pp. 21786–21791, 2002.
- [150] M. Gomelsky and G. Klug, "Bluf: a novel fad-binding domain involved in sensory transduction in microorganisms," *Trends Biochem. Sci.*, vol. 27, no. 10, pp. 497–500, 2002.
- [151] S. J. Davis, A. V. Vener, and R. D. Vierstra, "Bacteriophytochromes: Phytochrome-like photoreceptors from nonphotosynthetic eubacteria," *Science*, vol. 286, no. 5449, pp. 2517–2520, 1999.
- [152] M. J. McBride, "Cytophaga-flavobacterium gliding motility," *J. Mol. Microbiol. Biotechnol.*, vol. 7, no. 1-2, pp. 63–71, 2004.
- [153] I. A. Kataeva, R. D. Seidel, A. Shah, L. T. West, X. L. Li, and L. G. Ljungdahl, "The fibronectin type 3-like repeat from the *Clostridium thermocellum* cellobiohydrolase cbha promotes hydrolysis of cellulose by modifying its surface," *Appl. Environ. Microbiol.*, vol. 68, no. 9, pp. 4292–4300, 2002.
- [154] E. N. Karlsson, M. A. Hachem, S. Ramchuran, H. Costa, O. Holst, A. F. Svenningsen, and G. O. Hreggvidsson, "The modular xylanase xyn10a from *Rhodothermus marinus* is cell-attached, and its c-terminal domain has several putative homologues among cell-attached proteins within the phylum bacteroidetes," *FEMS Microbiol. Lett.*, vol. 241, no. 2, pp. 233–242, 2004.
- [155] L. Hall-Stoodley, J. W. Costerton, and P. Stoodley, "Bacterial biofilms: From the natural environment to infectious diseases," *Nat. Rev. Microbiol.*, vol. 2, no. 2, pp. 95–108, 2004.
- [156] C. M. Waters and B. L. Bassler, "Quorum sensing: Cell-to-cell communication in bacteria," *Annu. Rev. Cell Dev. Biol.*, vol. 21, pp. 319–346, 2005.
- [157] C. Mougel and I. B. Zhulin, "Chase: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants," *Trends Biochem. Sci.*, vol. 26, no. 10, pp. 582–584, 2001.
- [158] A. N. Nikolskaya and M. Y. Galperin, "A novel type of conserved dna-binding domain in the transcriptional regulators of the *algr/agra/lytr* family," *Nucleic Acids Res.*, vol. 30, no. 11, pp. 2453–2459, 2002.
- [159] A. W. Decho, "Microbial exopolymer secretions in ocean environments - their role(s) in food webs and marine processes," *Oceanography and Marine Biology: Annual Reviews*, vol. 28, pp. 73–153, 1990.
- [160] H. Ploug, M. Kühl, B. Buchholz-Cleven, and B. Joergensen, "Anoxic aggregates - an ephemeral phenomenon in the pelagic environment," *Aquat. Microb. Ecol.*, vol. 13, pp. 285–294, 1997.
- [161] O. Raize, Y. Argaman, and S. Yannai, "Mechanisms of biosorption of different heavy metals by brown marine macroalgae," *Biotechnol. Bioeng.*, vol. 87, no. 4, pp. 451–458, 2004.
- [162] M. Kube, A. Beck, S. H. Zinder, H. Kuhl, R. Reinhardt, and L. Adrian, "Genome sequence of the chlorinated compound-respiring bacterium *dehalococcoides* species strain *cbdb1*," *Nat. Biotechnol.*, vol. 23, no. 10, pp. 1269–1273, 2005.

- [163] T. Lowe and S. Eddy, "trnscan-se: a program for improved detection of transfer rna genes in genomic sequence," *Nucleic Acids Res.*, vol. 25, no. 5, pp. 955–964, 1997.
- [164] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: annotating non-coding rnas in complete genomes," *Nucleic Acids Res.*, vol. 33, pp. D121–D124, 2005.
- [165] H. Nielsen, S. Brunak, and G. von Heijne, "Machine learning approaches for the prediction of signal peptides and other protein sorting signals," *Protein Eng.*, vol. 12, pp. 3–9, 1999.
- [166] M. S. Rappe, K. Vergin, and S. J. Giovannoni, "Phylogenetic comparisons of a coastal bacterioplankton community with its counterparts in open ocean and freshwater systems," *FEMS Microbiol. Ecol.*, vol. 33, no. 3, pp. 219–232, 2000.
- [167] H. Eilers, J. Pernthaler, F. O. Glöckner, and R. Amann, "Culturability and in situ abundance of pelagic bacteria from the north sea," *Appl. Environ. Microbiol.*, vol. 66, no. 7, pp. 3044–3051, 2000.
- [168] K. M. Kelly and A. Y. Chistoserdov, "Phylogenetic analysis of the succession of bacterial communities in the great south bay (long island)," *FEMS Microbiol. Ecol.*, vol. 35, no. 1, pp. 85–95, 2001.
- [169] B. Crump, E. Armbrust, and J. Baross, "Phylogenetic analysis of particle-attached and free-living bacterial communities in the columbia river, its estuary, and the adjacent coastal ocean," *Appl. Environ. Microbiol.*, vol. 65, no. 7, pp. 3192–3204, 1999.
- [170] O. Beja, M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong, "Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage," *Environmental Microbiology*, vol. 2, no. 5, pp. 516–529, 2000.
- [171] S. G. Acinas, J. Antón, and F. Rodríguez-Valera, "Diversity of free-living and attached bacteria in offshore western mediterranean waters as depicted by analysis of genes encoding 16s rrna," *Appl. Environ. Microbiol.*, vol. 65, no. 2, pp. 514–522, 1999.
- [172] H. Schäfer, L. Bernard, C. Courties, P. Lebaron, P. Servais, R. Pukall, E. Stackebrandt, M. Troussellier, T. Guindulain, J. Vives-Rego, and G. Muyzer, "Microbial community dynamics in mediterranean nutrient-enriched seawater mesocosms: changes in the genetic diversity of bacterial populations," *FEMS Microbiol. Ecol.*, vol. 34, no. 3, pp. 243–253, 2001.
- [173] W. Ludwig, O. Strunk, S. Klugbauer, N. Klugbauer, M. Weizenegger, J. Neumeier, M. Bachleitner, and K.-H. Schleifer, "Bacterial phylogeny based on comparative sequence analysis," *Electrophoresis*, vol. 19, pp. 554–568, 1998.
- [174] S. A. Connon and S. J. Giovannoni, "High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates," *Appl. Environ. Microbiol.*, vol. 68, no. 8, pp. 3878–3885, 2002.
- [175] J.-C. Cho and S. J. Giovannoni, "Cultivation and growth characteristics of a diverse group of oligotrophic marine gammaproteobacteria," *Appl. Environ. Microbiol.*, vol. 70, no. 1, pp. 432–440, 2004.
- [176] R. Brinkmeyer, K. Knittel, J. Jurgens, H. Weyland, R. Amann, and E. Helmke, "Diversity and structure of bacterial communities in arctic versus antarctic pack ice," *Appl. Environ. Microbiol.*, vol. 69, no. 11, pp. 6610–6619, 2003.
- [177] H. Agogue, E. O. Casamayor, M. Bourrain, I. Obernosterer, F. Joux, G. J. Herndl, and P. Lebaron, "A survey on bacteria inhabiting the sea surface microlayer of coastal ecosystems," *FEMS Microbiol. Ecol.*, vol. 54, no. 2, pp. 269–280, 2005.
- [178] T. Maeda, K. Hayakawa, M. You, M. Sasaki, Y. Yamaji, M. Furushita, and T. Shiba, "Characteristics of nonylphenol polyethoxylate-degrading bacteria isolated from coastal sediments," *Microbes Environ.*, vol. 20, no. 4, pp. 253–257, 2005.
- [179] A. Pernthaler and J. Pernthaler, "Diurnal variation of cell proliferation in three bacterial taxa from coastal north sea waters," *Appl. Environ. Microbiol.*, vol. 71, no. 8, pp. 4638–4644, 2005.
- [180] C. E. Bauer, J. J. Buggy, Z. Yang, and B. L. Marrs, "The superoperonal organization of genes for pigment biosynthesis and reaction center proteins is a conserved feature in rhodobacter capsulatus: analysis of overlapping bcbh and puha transcripts," *Molecular and General Genetics MGG*, vol. 228, no. 3, pp. 433–444, 1991.



- [181] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, no. 1, p. 163, 2004.
- [182] O. Beja, M. T. Suzuki, J. F. Heidelberg, W. C. Nelson, C. M. Preston, T. Hamada, J. A. Eisen, C. M. Fraser, and E. F. DeLong, "Unsuspected diversity among marine aerobic anoxygenic phototrophs," *Nature*, vol. 415, no. 6872, pp. 630–633, 2002.
- [183] N. Yutin and O. Beja, "Putative novel photosynthetic reaction centre organizations in marine aerobic anoxygenic photosynthetic bacteria: insights from metagenomics and environmental genomics," *Environ. Microbiol.*, vol. 7, no. 12, pp. 2027–2033, 2005.
- [184] T. Suyama, T. Shigematsu, S. Takaichi, Y. Nodasaka, S. Fujikawa, H. Hosoya, Y. Tokiwa, T. Kanagawa, and S. Hanada, "Roseateles depolymerans gen. nov., sp. nov., a new bacteriochlorophyll a-containing obligate aerobe belonging to the beta-subclass of the proteobacteria," *Int. J. Syst. Bacteriol.*, vol. 49, pp. 449–457, 1999.
- [185] V. Yurkov and J. T. Beatty, "Isolation of aerobic anoxygenic photosynthetic bacteria from black smoker plume waters of the Juan de Fuca ridge in the Pacific Ocean," *Appl. Environ. Microbiol.*, vol. 64, no. 1, pp. 337–341, 1998.
- [186] J. A. Breznak, C. J. Potrikus, N. Pfennig, and J. C. Ensign, "Viability and endogenous substrates used during starvation survival of *Rhodospirillum rubrum*," *J. Bacteriol.*, vol. 134, no. 2, pp. 381–388, 1978.
- [187] S. Braatsch and G. Klug, "Blue light perception in bacteria," *Photosynth. Res.*, vol. 79, no. 1, pp. 45–57, 2004.
- [188] O. Preisig, D. Anthamatten, and H. Henneke, "Genes for a microaerobically induced oxidase complex in *Bradyrhizobium japonicum* are essential for a nitrogen-fixing endosymbiosis," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 90, no. 8, pp. 3309–3313, 1993.
- [189] J.-I. Oh and S. Kaplan, "Oxygen adaptation. the role of the ccoQ subunit of the cbb3 cytochrome c oxidase of *Rhodobacter sphaeroides*," *J. Biol. Chem.*, vol. 277, no. 18, pp. 16220–16228, 2002.
- [190] C. Alonso and J. Pernthaler, "Incorporation of glucose under anoxic conditions by bacterioplankton from coastal North Sea surface waters," *Appl. Environ. Microbiol.*, vol. 71, no. 4, pp. 1709–1716, 2005.
- [191] M. Krehenbrink, F.-B. Oppermann-Sanio, and A. Steinbüchel, "Evaluation of non-cyanobacterial genome sequences for occurrence of genes encoding proteins homologous to cyanophycin synthetase and cloning of an active cyanophycin synthetase from *Acinetobacter* sp. strain DSM 587," *Arch. Microbiol.*, vol. 177, no. 5, pp. 371–380, 2002.
- [192] Y. Elbahloul, M. Krehenbrink, R. Reichelt, and A. Steinbüchel, "Physiological conditions conducive to high cyanophycin content in biomass of *Acinetobacter calcoaceticus* strain adp1," *Appl. Environ. Microbiol.*, vol. 71, no. 2, pp. 858–866, 2005.
- [193] J. S. Mattick, "Type IV pili and twitching motility," *Annu. Rev. Microbiol.*, vol. 56, no. 1, pp. 289–314, 2002.
- [194] C. G. Friedrich, F. Bardischewsky, D. Rother, A. Quentmeier, and J. Fischer, "Prokaryotic sulfur oxidation," *Curr. Opin. Microbiol.*, vol. 8, no. 3, pp. 253–259, 2005.
- [195] F. Bardischewsky, J. Fischer, H. Bettina, and C. G. Friedrich, "SoxV transfers electrons to the periplasm of *Paracoccus pantotrophus* – an essential reaction for chemotrophic sulfur oxidation," *Microbiology*, vol. 152, p. 465–472, 2006.
- [196] C. Appia-Ayme and B. C. Berks, "SoxV, an orthologue of the ccdA disulfide transporter, is involved in thiosulfate oxidation in *Rhodovulum sulfidophilum* and reduces the periplasmic thioredoxin soxW," *Biochem. Biophys. Res. Commun.*, vol. 296, no. 3, pp. 737–741, 2002.
- [197] F. Bardischewsky and C. G. Friedrich, "The shxvW locus is essential for oxidation of inorganic sulfur and molecular hydrogen by *Paracoccus pantotrophus* gb17: a novel function for lithotrophy," *FEMS Microbiol. Lett.*, vol. 202, no. 2, pp. 215–220, 2001.
- [198] J. M. Gonzalez, R. Simo, R. Massana, J. S. Covert, E. O. Casamayor, C. Pedros-Alio, and M. A. Moran, "Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom," *Appl. Environ. Microbiol.*, vol. 66, no. 10, pp. 4237–4246, 2000.

- [199] M. Zubkov, B. Fuchs, S. Archer, R. Kiene, R. Amann, and P. Burkill, "Linking the composition of bacterioplankton to rapid turnover of dissolved dimethylsulphoniopropionate in an algal bloom in the north sea," *Environ. Microbiol.*, vol. 3, no. 5, pp. 304–311, 2001.
- [200] H. Ploug, "Small-scale oxygen fluxes and remineralization in sinking aggregates," *Limnol. Oceanogr.*, vol. 46, no. 7, p. 1624–1631, 2001.
- [201] D. de Beer, F. Wenzhoefer, T. G. Ferdelman, S. E. Boehme, M. Huettel, J. E. E. van Beusekom, M. E. Boettcher, N. Musat, and N. Dubilier, "Transport and mineralization rates in north sea sandy intertidal sediments, sylt-romo basin, wadden sea," *Limnol. Oceanogr.*, vol. 50, no. 1, pp. 113–127, 2005.
- [202] F. Azam and R. A. Long, "Oceanography: Sea snow microcosms," *Nature*, vol. 414, no. 6863, pp. 495–498, 2001.
- [203] H.-P. Grossart, T. Brinkhoff, T. Martens, C. Duerselen, G. Liebezeit, and M. Simon, "Tidal dynamics of dissolved and particulate matter and bacteria in a tidal flat ecosystem in spring and fall," *Limnol. Oceanogr.*, vol. 49, no. 6, pp. 2212–2222, 2004.
- [204] E. Ramalhosa, S. Segade, E. Pereira, C. Vale, and A. Duarte, "Mercury cycling between the water column and surface sediments in a contaminated area.," *Water Res.*, vol. 40, no. 15, pp. 2893–2900, 2006.
- [205] P. Divis, M. Leermakers, H. Docekalová, and Y. Gao, "Mercury depth profiles in river and marine sediments measured by the diffusive gradients in thin films technique with two different specific resins.," *Anal. Bioanal. Chem.*, vol. 382, no. 7, pp. 1715–1719, 2005.
- [206] O. Beja, L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong, "Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea," *Science*, vol. 289, no. 5486, pp. 1902–1906, 2000.
- [207] Z. S. Kolber, C. L. Van Dover, R. A. Niederman, and P. G. Falkowski, "Bacterial photosynthesis in surface waters of the open ocean," *Nature*, vol. 407, no. 6801, pp. 177–179, 2000.
- [208] Z. S. Kolber, F. G. Plumley, A. S. Lang, J. T. Beatty, R. E. Blankenship, C. L. VanDover, C. Vetriani, M. Koblizek, C. Rathgeber, and P. G. Falkowski, "Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean," *Science*, vol. 292, no. 5526, pp. 2492–2495, 2001.
- [209] M. T. Cottrell, A. Mannino, and D. L. Kirchman, "Aerobic anoxygenic phototrophic bacteria in the mid-atlantic bight and the north pacific gyre," *Appl. Environ. Microbiol.*, vol. 72, no. 1, pp. 557–564, 2006.
- [210] R. Goericke, "Bacteriochlorophyll a in the ocean: Is anoxygenic bacterial photosynthesis important?," *Limnol. Oceanogr.*, vol. 47, no. 1, pp. 290–295, 2002.
- [211] M. Schwalbach and J. A. Fuhrman, "Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative pcr," *Limnol. Oceanogr.*, vol. 50, no. 2, pp. 620–628, 2005.
- [212] M. Sieracki, I. Gilg, E. Thier, N. Poulton, and R. Goericke, "Distribution of planktonic aerobic anoxygenic photoheterotrophic bacteria in the northwest atlantic," *Limnol. Oceanogr.*, vol. 51, no. 1, pp. 38–46, 2006.
- [213] T. Shiba, "Roseobacter litoralis new-genus new-species and roseobacter denitrificans new-species aerobic pink-pigmented bacteria which contain bacteriochlorophyll a," *System. Appl. Microbiol.*, vol. 14, no. 2, pp. 140–145, 1991.
- [214] N. Selje, M. Simon, and T. Brinkhoff, "A newly discovered roseobacter cluster in temperate and polar oceans," *Nature*, vol. 427, no. 6973, pp. 445–448, 2004.
- [215] S. M. D. Goldberg, J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y.-H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier, and J. C. Venter, "A sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 30, pp. 11240–11245, 2006.

- [216] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, "A whole-genome assembly of drosophila," *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000.
- [217] P. Rice, I. Longden, and A. Bleasby, "Emboss: The european molecular biology open software suite," *Trends Genet.*, vol. 16, no. 6, pp. 276–277, 2000.
- [218] P. F. Hallin, T. T. Binnewies, and D. W. Ussery, "Genome update: chromosome atlases," *Microbiology*, vol. 150, no. 10, pp. 3091–3093, 2004.
- [219] R. Gibson and D. R. Smith, "Genome visualization made fast and simple," *Bioinformatics*, vol. 19, no. 11, pp. 1449–1450, 2003.
- [220] M. Borodovsky and J. McIninch, "Genemark: parallel gene recognition for both dna strands," *Comput. Chem.*, vol. 17, no. 19, pp. 123–133, 1993.
- [221] A. Lukashin and M. Borodovsky, "Genemark.hmm: new solutions for gene finding.," *Nucleic Acids Res.*, vol. 26, no. 4, pp. 1107–1115, 1998.
- [222] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, "Improved microbial gene identification with GLIMMER," *Nucleic Acids Res.*, vol. 27, no. 23, pp. 4636–4641, 1999.
- [223] R. Tatusov, N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, B. S. Rao, S. Smirnov, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin, and D. Natale, "The cog database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, no. 1, p. 41, 2003.
- [224] P. Gerhardt, *Methods for general and molecular bacteriology*. Washington D.C.: American Society for Microbiology, 1994.
- [225] N. P. Revsbech, "An oxygen microelectrode with a guard cathode.," *Limnol. Oceanogr.*, vol. 34, pp. 474–478, 1989.
- [226] S. Wright and S. Jeffrey, "High-resolution hplc system for chlorophylls and carotenoids of marine phytoplankton.," in *Phytoplankton pigments in oceanography: guidelines to modern methods* (S. Jeffrey, R. Mantoura, and S. Wright, eds.), pp. 327 – 342, Paris: UNESCO, 1997.
- [227] A. Hiraishi, H. Kuraishi, and K. Kawahara, "Emendation of the description of blastomonas natoria (sly 1985) sly and cahill 1997 as an aerobic photosynthetic bacterium and reclassification of erythromonas ursincola yurkov et al. 1997 as blastomonas ursincola comb. nov.," *Int. J. Syst. Evol. Microbiol.*, vol. 50, no. 3, pp. 1113–1118, 2000.
- [228] D. Field and N. Kyrpides, "The positive role of the ecological community in the genomic revolution.," *Microb Ecol*, vol. 53, pp. 507–511, Apr 2007.
- [229] H. Schlesner, C. Rensmann, B. J. Tindall, D. Gade, R. Rabus, S. Pfeiffer, and P. Hirsch, "Taxonomic heterogeneity within the planctomycetales as derived by dna-dna hybridization, description of rhodopirellula baltica gen. nov., sp. nov., transfer of pirellula marina to the genus blastopirellula gen. nov. as blastopirellula marina comb. nov. and emended description of the genus pirellula," *Int. J. Syst. Evol. Microbiol.*, vol. 54, no. Pt 5, pp. 1567–80, 2004.
- [230] A. Chatzinotas, R. Sandaa, W. Schonhuber, R. Amann, F. Daae, V. Torsvik, J. Zeyer, and D. Hahn, "Analysis of broad-scale differences in microbial community composition of two pristine forest soils," *System. Appl. Microbiol.*, vol. 21, no. 4), pp. 579–587, 1998.
- [231] A. Neef, R. Amann, H. Schlesner, and K. H. Schleifer, "Monitoring a widespread bacterial group: in situ detection of planctomycetes with 16s rrna-targeted probes," *Microbiology*, vol. 144 ( Pt 12), pp. 3257–66, 1998.
- [232] K. L. Vergin, E. Urbach, J. L. Stein, E. F. DeLong, B. D. Lanoil, and S. J. Giovannoni, "Screening of a fosmid library of marine environmental genomic dna fragments reveals four clones related to members of the order planctomycetales," *Appl. Environ. Microbiol.*, vol. 64, no. 8, pp. 3075–3078, 1998.
- [233] S. J. Giovannoni, E. Schabtach, and R. W. Castenholz, "Isosphaera pallida, gen. and comb. nov., a gliding, budding eubacterium from hot springs," *Arch. Microbiol.*, vol. 147, pp. 276–284, April 1987.

- [234] S. Pimentel-Elardo, M. Wehrl, A. B. Friedrich, P. R. Jensen, and U. Hentschel, "Isolation of planctomycetes from aplysina sponges," *Aquat. Microb. Ecol.*, vol. 33, no. 3, pp. 239–245, 2003.
- [235] J. A. Fuerst, H. G. Gwilliam, M. Lindsay, A. Lichanska, C. Belcher, J. E. Vickers, and P. Hugenholz, "Isolation and molecular identification of planctomycete bacteria from postlarvae of the giant tiger prawn, *penaeus monodon*," *Appl. Environ. Microbiol.*, vol. 63, no. 1, pp. 254–62, 1997.
- [236] H. König, H. Schlesner, and P. Hirsch, "Cell wall studies on budding bacteria of the planctomyces pasteuria group and on a prosthecomicrobium sp.," *Arch. Microbiol.*, vol. 138, no. 3, pp. 200–205, 1984.
- [237] W. Liesack, H. König, H. Schlesner, and P. Hirsch, "Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the *pirellula*/planctomyces group," *Arch. Microbiol.*, vol. 145, pp. 361–366, 1986.
- [238] A. Fuerst, "Intracellular compartmentation in planctomycetes," *Annu. Rev. Microbiol.*, vol. 59, pp. 299–328, 2005.
- [239] B. L. Tekniepe, J. M. Schmidt, and M. P. Starr, "Life-cycle of a budding and appendaged bacterium belonging to morphotype-iv of the *blastocaulis*-planctomyces group," *Curr. Microbiol.*, vol. 5, no. 1, pp. 1–6, 1981.
- [240] J. A. Fuerst, "The planctomycetes: emerging models for microbial ecology, evolution and cell biology," *Microbiology*, vol. 141 ( Pt 7), pp. 1493–506, 1995.
- [241] C. Jacobs-Wagner, "Regulatory proteins with a sense of direction: cell cycle signalling network in *Caulobacter*," *Mol. Microbiol.*, vol. 51, no. 1, pp. 7–13, 2004.
- [242] M. Y. Galperin and E. V. Koonin, "conserved hypothetical' proteins: prioritization of targets for experimental study," *Nucleic Acids Res.*, vol. 32, no. 18, pp. 5452–5463, 2004.
- [243] R. Rabus, D. Gade, R. Helbig, M. Bauer, F. O. Glockner, M. Kube, H. Schlesner, R. Reinhardt, and R. Amann, "Analysis of n-acetylglucosamine metabolism in the marine bacterium *pirellula* sp. strain 1 by a proteomic approach," *Proteomics*, vol. 2, no. 6, pp. 649–55, 2002.
- [244] D. Gade, T. Stuhmann, R. Reinhardt, and R. Rabus, "Growth phase dependent regulation of protein composition in *rhodopirellula baltica*," *Environ. Microbiol.*, vol. 7, no. 8, pp. 1074–1084, 2005.
- [245] D. Gade, J. Gobom, and R. Rabus, "Proteomic analysis of carbohydrate catabolism and regulation in the marine bacterium *rhodopirellula baltica*," *Proteomics*, vol. 5, no. 14, pp. 3672–3683, 2005.
- [246] D. Gade, J. Thiermann, D. Markowsky, and R. Rabus, "Evaluation of two-dimensional difference gel electrophoresis for protein profiling. soluble proteins of the marine bacterium *pirellula* sp. strain 1," *J. Mol. Microbiol. Biotechnol.*, vol. 5, no. 4, pp. 240–51, 2003.
- [247] C. X. Hieu, B. Voigt, D. Albrecht, D. Becher, T. Lombardot, F. O. Glöckner, R. Amann, M. Hecker, and T. Schweder, "Detailed proteome analysis of growing cells of the planctomycete *rhodopirellula baltica* sh1(t)," *Proteomics*, vol. 8, no. 8, pp. 1608–1623, 2008.
- [248] S. R. Wallner, M. Bauer, C. Würdemann, P. Wecker, F. O. Glöckner, and K. Faber, "Highly enantioselective sec-alkylsulfatase activity of the marine planctomycete *rhodopirellula baltica* shows retention of configuration," *Angew. Chem. Int. Ed.*, vol. 44, pp. 2–4, 2005.
- [249] J. Dabin, M. Jam, M. Czjzek, and G. Michel, "Expression, purification, crystallization and preliminary x-ray analysis of the polysaccharide lyase rb5312 from the marine planctomycete *rhodopirellula baltica*," *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 64, no. Pt 3, pp. 224–7, 2008.
- [250] H. Gao, Y. Wang, X. Liu, T. Yan, L. Wu, E. Alm, A. Arkin, D. K. Thompson, and J. Zhou, "Global Transcriptome Analysis of the Heat Shock Response of *Shewanella oneidensis*," *J. Bacteriol.*, vol. 186, no. 22, pp. 7796–7803, 2004.
- [251] H. Gao, Z. K. Yang, L. Wu, D. K. Thompson, and J. Zhou, "Global Transcriptome Analysis of the Cold Shock Response of *Shewanella oneidensis* MR-1 and Mutational Analysis of Its Classical Cold Shock Proteins," *J. Bacteriol.*, vol. 188, no. 12, pp. 4560–4569, 2006.
- [252] A. Aspedon, K. Palmer, and M. Whiteley, "Microarray analysis of the osmotic stress response in *pseudomonas aeruginosa*," *J. Bacteriol.*, vol. 188, no. 7, pp. 2721–2725, 2006.

- [253] A. Mukhopadhyay, Z. He, E. J. Alm, A. P. Arkin, E. E. Baidoo, S. C. Borglin, W. Chen, T. C. Hazen, Q. He, H.-Y. Holman, K. Huang, R. Huang, D. C. Joyner, N. Katz, M. Keller, P. Oeller, A. Redding, J. Sun, J. Wall, J. Wei, Z. Yang, H.-C. Yen, J. Zhou, and J. D. Keasling, "Salt stress in *Desulfovibrio vulgaris* hildenborough: an integrated genomics approach.," *J. Bacteriol.*, vol. 188, pp. 4068–4078, Jun 2006.
- [254] T. Koide, R. Z. N. Vencio, and S. L. Gomes, "Global Gene Expression Analysis of the Heat Shock Response in the Phytopathogen *Xylella fastidiosa*," *J. Bacteriol.*, vol. 188, no. 16, pp. 5821–5830, 2006.
- [255] Y. Kanesaki, I. Suzuki, S. I. Allakhverdiev, K. Mikami, and N. Murata, "Salt stress and hyperosmotic stress regulate the expression of different sets of genes in *Synechocystis* sp. pcc 6803," *Biochem. Biophys. Res. Commun.*, vol. 290, no. 1, pp. 339 – 348, 2002.
- [256] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes," *Mol. Biol. Cell*, vol. 11, no. 12, pp. 4241–4257, 2000.
- [257] S. Tachdjian and R. M. Kelly, "Dynamic metabolic adjustments and genome plasticity are implicated in the heat shock response of the extremely thermoacidophilic archaeon *Sulfolobus solfataricus*," *J. Bacteriol.*, vol. 188, pp. 4553–4559, Jun 2006.
- [258] H. Schlesner, "The development of media suitable for the microorganisms morphologically resembling planctomyces spp., *Pirellula* spp., and other planctomycetales from various aquatic habitats using dilute media," *System. Appl. Microbiol.*, vol. 17, pp. 135–145, 1994.
- [259] P. G. Jones, M. Cashel, G. Glaser, and F. C. Neidhardt, "Function of a relaxed-like state following temperature downshifts in *Escherichia coli*," *J. Bacteriol.*, vol. 174, pp. 3903–3914, Jun 1992.
- [260] J. A. Coker, P. DasSarma, J. Kumar, J. A. Müller, and S. DasSarma, "Transcriptional profiling of the model archaeon *Halobacterium* sp. nrc-1: responses to changes in salinity and temperature.," *Saline Syst.*, vol. 3, p. 6, 2007.
- [261] S. M. Sowell, A. D. Norbeck, M. S. Lipton, C. D. Nicora, S. J. Callister, R. D. Smith, D. F. Barofsky, and S. J. Giovannoni, "Proteomic analysis of stationary phase in the marine bacterium "*Candidatus Pelagibacter ubique*" .," *Appl. Environ. Microbiol.*, vol. 74, pp. 4091–4100, Jul 2008.
- [262] S. Derzelle, B. Hallet, T. Ferain, J. Delcour, and P. Hols, "Improved Adaptation to Cold-Shock, Stationary-Phase, and Freezing Stresses in *Lactobacillus plantarum* Overproducing Cold-Shock Proteins," *Appl. Environ. Microbiol.*, vol. 69, no. 7, pp. 4285–4290, 2003.
- [263] S. Phadtare, "Recent developments in bacterial cold-shock response.," *Curr. Issues. Mol. Biol.*, vol. 6, pp. 125–136, Jul 2004.
- [264] J. Ott, *Meereskunde*. Stuttgart: Ulmer Verlag, 1996.
- [265] G. Rheinheimer, *Meereskunde der Ostsee*. Berlin, Heidelberg, New York: Springer-Verlag, 2 ed., 1996.
- [266] B. Kempf and E. Bremer, "Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments.," *Arch. Microbiol.*, vol. 170, pp. 319–330, Oct 1998.
- [267] M. Roesser and V. Müller, "Osmoadaptation in bacteria and archaea: common principles and differences.," *Environ. Microbiol.*, vol. 3, pp. 743–754, Dec 2001.
- [268] H. J. Kunte, "Osmoregulation in bacteria: Compatible solute accumulation and osmosensing," *Environ. Chem.*, vol. 3, no. 2, pp. 94–99, 2006.
- [269] D. J. Studholme, J. A. Fuerst, and A. Bateman, "Novel protein domains and motifs in the marine planctomycete *Rhodopirellula baltica*," *FEMS Microbiol. Lett.*, vol. 236, no. 2, pp. 333–340, 2004.
- [270] C. Vargas, M. Argandoña, M. Reina-Bueno, J. Rodríguez-Moya, C. Fernández-Aunión, and J. J. Nieto, "Unravelling the adaptation responses to osmotic and temperature stress in chromohalobacter *Salexigens*, a bacterium with broad salinity tolerance.," *Saline Syst.*, vol. 4, p. 14, 2008.
- [271] C. D. Pivetti, M.-R. Yen, S. Miller, W. Busch, Y.-H. Tseng, I. R. Booth, and M. H. Saier, "Two families of mechanosensitive channel proteins.," *Microbiol. Mol. Biol. Rev.*, vol. 67, pp. 66–85, table of contents, Mar 2003.
- [272] T. Zeller and G. Klug, "Thioredoxins in bacteria: functions in oxidative stress response and regulation of thioredoxin genes.," *Naturwissenschaften*, vol. 93, pp. 259–266, Jun 2006.

- [273] P. M. Pereira, Q. He, A. V. Xavier, J. Zhou, I. A. C. Pereira, and R. O. Louro, "Transcriptional response of *Desulfovibrio hildenborough* to oxidative stress mimicking environmental conditions.," *Arch. Microbiol.*, vol. 189, pp. 451–461, May 2008.
- [274] D. J. Studholme and R. Dixon, "In silico analysis of the sigma(54)-dependent enhancer-binding proteins in *Pirellula* species strain 1," *FEMS Microbiol. Lett.*, vol. 230, no. 2, pp. 215–225, 2004.
- [275] A. Marchler-Bauer, J. B. Anderson, M. K. Derbyshire, C. DeWeese-Scott, N. R. Gonzales, M. Gwadz, L. Hao, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, D. Krylov, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, N. Thanki, R. A. Yamashita, J. J. Yin, D. Zhang, and S. H. Bryant, "Cdd: a conserved domain database for interactive domain family analysis.," *Nucleic Acids Res.*, vol. 35, pp. D237–D240, Jan 2007.
- [276] T. Lombardot, M. Bauer, H. Teeling, R. Amann, and F. O. Glöckner, "The transcriptional regulator pool of the marine bacterium *Rhodopirellula baltica* SH1<sup>T</sup> as revealed by whole genome comparisons," *FEMS Microbiol. Lett.*, vol. 242, no. 1, pp. 137–145, 2005.
- [277] K. Mikami, Y. Kanesaki, I. Suzuki, and N. Murata, "The histidine kinase hik33 perceives osmotic stress and cold stress in *Synechocystis* sp pcc 6803.," *Mol. Microbiol.*, vol. 46, pp. 905–915, Nov 2002.
- [278] R. A. VanBogelen and F. C. Neidhardt, "Ribosomes as sensors of heat and cold shock in *Escherichia coli*," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 87, pp. 5589–5593, Aug 1990.
- [279] H. A. Thieringer, P. G. Jones, and M. Inouye, "Cold shock and adaptation.," *Bioessays*, vol. 20, pp. 49–57, Jan 1998.
- [280] O. Reva and B. Tümmler, "Think big—giant genes in bacteria.," *Environ. Microbiol.*, vol. 10, pp. 768–777, Mar 2008.
- [281] A. Soukas, P. Cohen, N. D. Socci, and J. M. Friedman, "Leptin-specific patterns of gene expression in white adipose tissue.," *Genes Dev.*, vol. 14, pp. 963–980, Apr 2000.
- [282] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush, "Tm4: a free, open-source system for microarray data management and analysis," *BioTechniques*, vol. 34, no. 2, pp. 374–8, 2003.
- [283] C. Quast, "MicHanThi - Design and Implementation of a System for the Prediction of Gene Functions in Genome Annotation Projects," tech. rep., University of Bremen, 2006.
- [284] "The reannotated *Rhodopirellula baltica* genome [<http://gendb.mpi-bremen.de/gendb/BX119912>]."
- [285] M. Richter, T. Lombardot, I. Kostadinov, R. Kottmann, M. B. Duhaime, J. Peplies, and F. O. Glöckner, "Jcoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta) genomes," *BMC Bioinformatics*, vol. 9, p. 177, 2008.
- [286] T. Hulsen, J. de Vlieg, and W. Alkema, "BioVenn - a web application for the comparison and visualization of biological lists using area-proportional venn diagrams.," *BMC Genomics*, vol. 9, p. 488, 2008.
- [287] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Sobolova, M. Tomashevsky, and R. Edgar, "Ncbi geo: mining tens of millions of expression profiles—database and tools update.," *Nucleic Acids Res.*, vol. 35, pp. D760–D765, Jan 2007.
- [288] C. Woese, "Bacterial evolution," *Microbiol. Rev.*, vol. 51, no. 2, pp. 221–271, 1987.
- [289] G. J. Olsen, C. R. Woese, and R. Overbeek, "The winds of (evolutionary) change: Breathing new life into microbiology," *J. Bacteriol.*, vol. 176, no. 1, pp. 1–6, 1994.
- [290] C. Cary and P. Chisholm, "Report of a workshop on marine microbial genomics to develop recommendations for the national science foundation," tech. rep., Arlington, VA, 19–20 April 2004 2000.
- [291] E. F. DeLong and D. M. Karl, "Genomic perspectives in microbial oceanography.," *Nature*, vol. 437, pp. 336–342, Sep 2005.
- [292] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "Genbank," *Nucleic Acids Res.*, vol. 33, pp. D34–D38, 2005.

- [293] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy, "The pfam protein families database," *Nucleic Acids Res.*, vol. 32, pp. D138–D141, 2004.
- [294] A. Dufresne, M. Salanoubat, F. Partensky, F. Artiguenave, I. M. Axmann, V. Barbe, S. Duprat, M. Y. Galperin, E. V. Koonin, F. Le Gall, K. S. Makarova, M. Ostrowski, S. Oztas, C. Robert, I. B. Rogozin, D. J. Scanlan, N. T. de Marsac, J. Weissenbach, P. Wincker, Y. I. Wolf, and W. R. Hess, "Genome sequence of the cyanobacterium *prochlorococcus marinus* ss120, a nearly minimal oxyphototrophic genome," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 17, pp. 10020–10025, 2003.
- [295] G. Rocop, F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser, and S. W. Chisholm, "Genome divergence in two *prochlorococcus* ecotypes reflects oceanic niche differentiation," *Nature*, vol. 424, no. 6952, pp. 1042–1047, 2003.
- [296] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environ. Microbiol.*, vol. 6, no. 9, pp. 938–947, 2004.