Max Planck Institute
for Marine Microbiology

JACOBS UNIVERSITY
School of Engineering and Science

# Data integration in microbial genomics: *Contextualizing sequence data in aid of biological knowledge*

by

**Wolfgang Matthias Hankeln, Dipl. Inf., M.Sc.**

A thesis submitted in partial fulfillment of requirements for the degree of

Doctor Of Philosophy
in Bioinformatics

| | |
|---|---|
| Approved thesis committee: | Prof. Dr. Frank Oliver Glöckner (chair)<br>Max Planck Institute for Marine Microbiology<br>Jacobs University |
| | Prof. Dr. Peter Baumann<br>Jacobs University |
| | Dr. Wolfgang Ludwig<br>Technische Universität München |
| Date of defense: | 30 May 2011 |

# Statement of Sources

DECLARATION

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from published or unpublished scientific work has been cited in the text and listed in the references.

**Signature** **Date**

# List of abbreviations

**BLAST** Basic Local Alignment and Search Tool

**DIKW** ”data information knowledge wisdom hierarchy”

**DDBJ** DNA Data Bank of Japan

**DNA** deoxyribonucleic acid

**DUF** Domain of Unknown Function

**ENA** European Nucleotide Archive

**FOL** First Order Logic

**GOS** Global Ocean Survey

**GSC** Genomic Standards Consortium

**HMM** Hidden Markov Model

**INSDC** International Nucleotide Sequence Database Collaboration

**KEGG** Kyoto Encyclopedia of Genes and Genomes

**MarMic** Marine Microbiology

**MIMARKS** “Minimum Information about a MARKer Sequence”

**MIGS** “Minimum Information about a Genome Sequence”

**MIMS** ”Minimum Information about a Metagenome Sequence“

**MIxS** Minimum Information about any (x) Sequence

**NCBI** National Center for Biotechnology Information

**NGS** Next Generation Sequencing

**ORF** Open Reading Frame

**RNA** ribonucleic acid

**rRNA** ribosomal ribonucleic acid

**WOA** World Ocean Atlas

***“To me, bioinformatics has an inherent ’cool factor’ because it challenges our approaches to analytics, pushes the envelope on performance, and offers a compelling example for a web services-based industry.”***

**(James Gosling, O’Reilly Media, 2003)**

# Thesis abstract

Deoxyribonucleic acid (DNA) is the primary structure that carries the genetic information of organisms in genomes. The introduction of the first DNA sequencing methods in 1977 marked a major breakthrough in life sciences. Today, these methods are widely applied and grant insight into the 'blueprints' of organisms from all domains of life.

The analysis of environmental microbial sequence data is becoming increasingly important in times of global climate change, because microbes are central catalysts in nutrient cycles such as the carbon cycle that profoundly affects Earth's climate. Microbes perform almost all metabolic processes that are thermodynamically possible.

DNA sequencing is carried out around the globe and the resulting data is submitted to the public repositories of the International Nucleotide Sequence Database Collaboration (INSDC). Data in the INSDC is accumulating exponentially. This trend shows the need for efficient data processing strategies in order to gain knowledge out of this ever increasing amount of sequence data.

For this, it is important to annotate sequence data with as much contextual data as possible. Contextual data are data about the environmental context and the processing steps that were applied. These can range from data about the geographic location, sampling time, habitat, or about experimental procedures used to obtain the sequences up to video data recorded during sampling. Especially data about the geographic location (x, y, z) and the point in time (t), when samples are taken from the environment are essential. Comparability and interpretability are preserved. Ample analysis approaches become possible, when contextual and sequence data are integrated.

In this doctoral thesis, data integration is promoted in three ways: Firstly, through the development of contextual data capture, submission and integration tools. Secondly, through the development of standards for contextual data and thirdly, through demonstration of *in silico* hypothesis generation for a large metagenomic data set.

# Contents

CHAPTER 1

# INTRODUCTION

*“We’re drowning in information and starving for knowledge.”*
(Rutherford D. Rogers, NY Times, 1988)

In this introductory chapter, the basic concepts and a theoretical framework are presented to set the stage. The focus of this doctoral thesis is put into a broader context. Subsequently, the relevant bioinformatic approaches and problems are presented that are tackled by the research aims. The chapter finishes with an overview about the overall structure of the thesis.

It is the intention to guide a reader not familiar with the subject of this thesis towards an understanding that helps to fully appreciate the motivation of the research aims and the studies conducted.

## 1.1 From data to knowledge

*“Of our mundane and technical concepts information is currently one of the most important, most widely used and least understood.”*
[Floridi, 2003]

Information as well as data and knowledge are central terms in this thesis. They are defined in the following. This happens without claiming to offer the only philosophically valid definitions for these terms, but to establish a consistent vocabulary that will be used throughout this work.

Information scientists agree that on the lowest level there are signs that are defined in alphabets (see figure 1.1). These signs can be com-

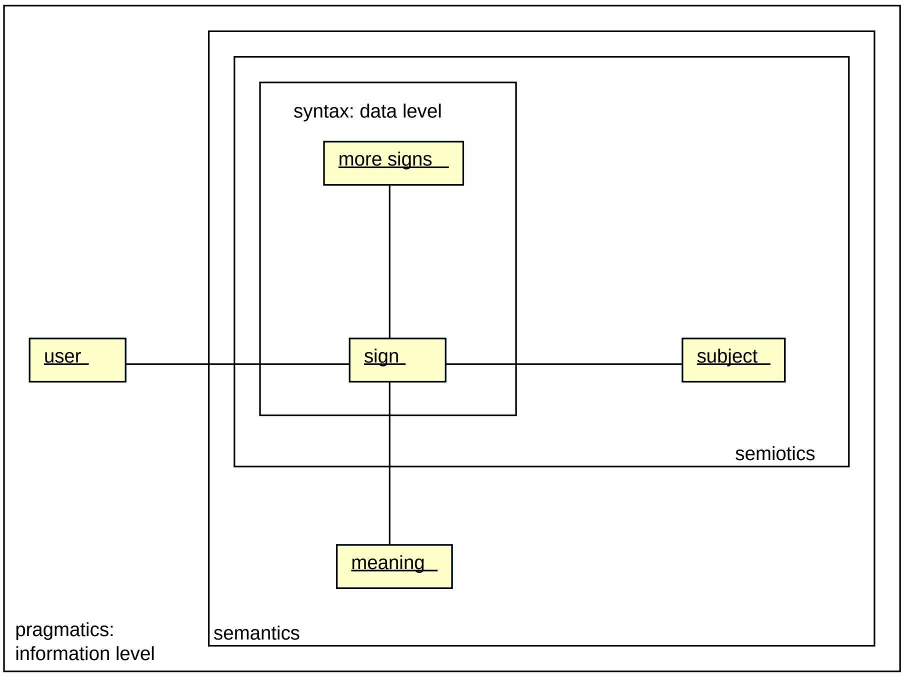


Figure 1.1: data and information, modified from: [Berthel, 1975]

**bined to form expressions, which are called data. Rules that describe how these signs are validly combined are called syntax.**

**Valid data expressions can often be combined to form even larger expressions. Whereas alphabets have a finite size, data expressions can in theory be infinite[1]. On the level of semantics, these data expressions become meaningful. When data are interpreted in their context by a user, they become information. Information can be defined as interpreted data. To transform data into information, it must be clear, to which subject these data are related to and in which context they were created. Thus, information is always subject-related and context-based [Rowley, 2007]. Information help to answer questions like "Who...?", "How much ...?", "How many ...?", " What ...?", " Where ...?" and "When ...?" [Ackoff, 1989]. Figure 1.2 shows an example. On the lowest level is an expression that consists of signs that are combined according to a syntax. The data expression alone is not meaningful, until this data is correctly interpreted to be an exchange rate. Once interpreted, the data expression becomes information.**

**Whereas data and information can be clearly distinguished, it is somewhat harder to differentiate information and knowledge. There is**

[1] In reality data has to be stored on a physical medium such as a harddrive and therefore can not be infinite in size [Beynon-Davies, 2002].

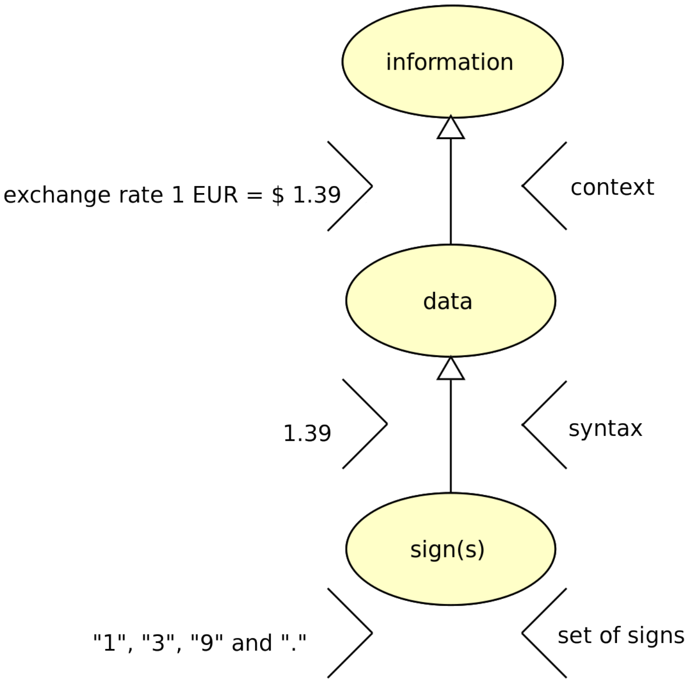


Figure 1.2: Exchange rate example, modified from: [Krcmar, 2009]

**still no consensus on the nature of knowledge among information scientists [Rowley, 2007]. For this work, knowledge is defined as condensed information that can be obtained through empirical and thorough analysis of data and information by applying reproducible methods[2]. Furthermore, knowledge aids decision making. One aspect of knowledge is that it is always bound to the individual and has to be learned. Once learned, knowledge becomes a part of the mind and is then called tacit or implicit knowledge. In contrast, explicit knowledge has been recorded in some form that allows an individual to transform this explicit knowledge into tacit knowledge again [P. Bocij and Hickie, 2005]. Science mainly focuses on explicit knowledge that can be empirically derived and reproduced in experiments and recorded in publications, books or databases.**

**In computer science, knowledge can be computed. When a real world domain has been mapped onto a computable formalism, knowledge can be derived from it. Computable formalisms are for example First Order Logic (FOL) or ontologies. FOL is a formal logical system which consists of a formal language and a set of inference rules that are**

[2] This definition is consistent with the definition of knowledge that is obtained by the application of scientific methods [Wilson, 1991].

**used to derive new expressions from one or more expressions called premises. An ontology is a *“formal, explicit specification of a shared conceptualisation“* [Gruber, 1993]. Ontologies provide a shared vocabulary, which can be used to model a domain. A domain is comprised of objects and/or concepts that exist and their properties and relations. Knowledge encoded in FOL or ontologies can be stored in so-called knowledge bases. *”In artificial intelligence, FOL is used ... to encode knowledge that then can be stored and used by a computer”* [Etchemendy, 1999]. Computers can then infer new knowledge from previously encoded knowledge. Inference is defined as *“a way to add new sentences to the knowledge base by deriving new sentences from old“* [Norvig, 2009]. An example in natural language is to infer from the premises “All men are mortal” and ”Socrates is a man” that “Socrates is mortal”[3]. The larger a knowledge base, the more applicable are computers to infer new knowledge. Computers are able to rapidly compute vast amounts of data and knowledge, while human minds become overstrained by sheer quantity.**

**Whereas knowledge can be computed, wisdom is a domain that is still reserved to humans exclusively, perhaps because it *”has more to do with human intuition, understanding, interpretation and actions, than with systems”* [Rowley, 2007].**

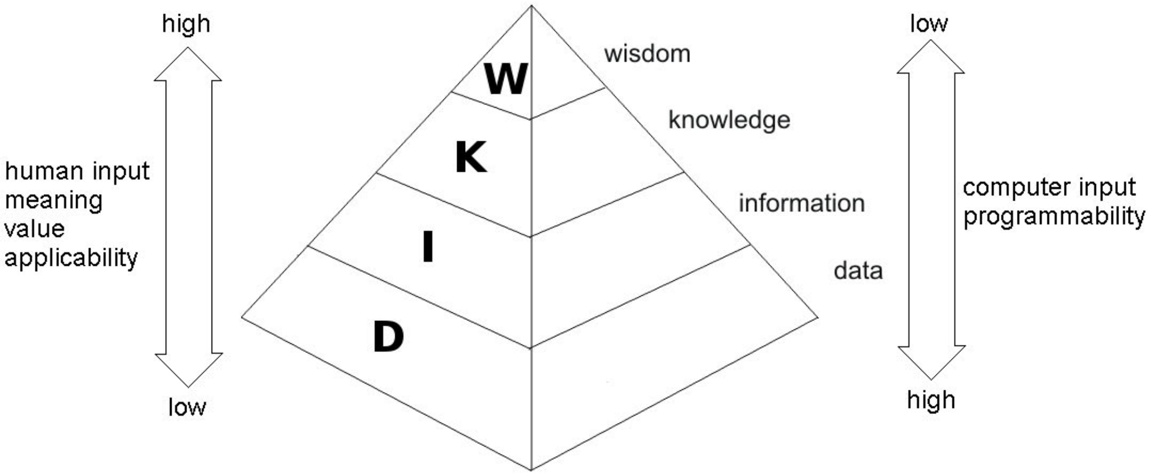


Figure 1.3: Knowledge pyramid, modified from: [Rowley, 2007]

**The ”data information knowledge wisdom hierarchy” (DIKW) or**

[3] Whereas natural language is often ambiguous, FOL and ontologies are not. The logic in these sentences is extracted and formalized to unambiguous expressions that can be computed [Etchemendy, 1999].

**”Knowledge Pyramid” shown in figure 1.3 is a model that is often used in the information and knowledge literature to provide a combined overview about the key terms and how they are related. The shape of the pyramid implies that there are a lot of data at the bottom of the pyramid. These data can be used to obtain information. On the next level there is fewer information that can be used to obtain knowledge. Finally, knowledge can be used to obtain wisdom.** ***“Each of the higher types in the hierarchy includes the categories that fall below it” [Ackoff, 1989].***

**Shown on the left and right of the pyramid are variables that change on each level of the hierarchy from “high“ to “low“. Data at the lowest level can be computed easily. They have a low meaning and require few human input. It is a tedious task that requires a lot of human input, to model a knowledge domain as an ontology that then can be computed. The higher the level in the pyramid, the lower is the computer input programmability and the more human input is required. At the top of the pyramid the meaning, applicability and value is the highest but the programmability is the lowest.**

**After the key terms and the knowledge pyramid have been explained, the field of microbiology is briefly introduced. Subsequently, the relevant bioinformatic approaches are explained, and how they can be applied to come from data to knowledge in marine microbial genomics.**

## 1.2 Relevance of microbiology

**Our planet is inhabited by an unseen majority of microorganisms[4] [Whitman et al., 1998] that are central catalysts in the global cycling of nutrients. Microorganisms form two separate and extremely diverse domains[5] of life that are collectively termed the prokaryotes. It is estimated that microbes contribute between 50% and 90% to global primary production [Falkowski et al., 1998]. Almost all life on earth directly or indirectly relies on this production of organic compounds from atmospheric or aquatic carbon dioxide [Field et al., 1998]. Mi-**

[4] The total number of prokaryotes is estimated to be $4 - 6 * 10^{30}$ on Earth [Whitman et al., 1998].

[5] The domains are *Archaea* and *Bacteria*.

**croorganisms thrive in all kinds of habitats. They are found almost everywhere on Earth: in arctic ice, acid mine drainages, air, soil and water up to extremely hot hydrothermal vents in the deep sea [Rappé and Giovannoni, 2003]. Furthermore, microorganisms perform almost all metabolic processes that are thermodynamically possible [Henry et al., 2007], many of which are likely to be of interest to industry[6]. Especially in the oceans that cover two thirds of Earth's surface [Rahmstorf, 2002], microorganisms are involved in the carbon and nitrogen cycle[7] that profoundly affect Earth's climate [Arrigo, 2005, Zehr and Kudela, 2011]. In times of global climate change, knowledge about the complex interplay of marine microbes in the biosphere is needed to understand, predict, and potentially address global changes [Glöckner and Joint, 2010].**

**It is therefore interesting to know "Who is out there?", "How many of which kind?" and "What are they doing?". These are the central questions of marine molecular ecology, which is a subdiscipline in the field of marine microbiology [Amann, 2000]. Though these three questions sound trivial, it is hard to answer them. Microorganisms are so small in size and large in number that it is hard to count and differentiate them based on optical features. Microorganisms are mostly unicellular. They replicate by cell division. The smallest nano bacteria are around 0.2 micrometer, the largest bacterium, namely *Thiomargerita namibiensis* is 750 micrometer in diameter [Schulz and Jorgensen, 2001]. It is estimated that one milliliter water contains roughly $5 * 10^5$ cells and a gram of soil contains roughly $2 * 10^9$ cells [Whitman et al., 1998]. Due to their small size, microbes are transported and globally dispersed, when water evaporates to the atmosphere, with water currents in the ocean and attached to host organisms. This has first been observed in 1927 and led to the hypothesis that *"everything is everywhere, but the environment selects*[8]*"* [Meehan and Baas-Becking, 1927]. This hypothesis is still debated among the life science community.**

[6] Together with the fact that microbes also have and impact to the health of higher, multicellular, organisms, this points out their relevance to the medical and biotechnological field.

[7] The impact is mainly due to greenhouse gases like carbon dioxide, methane and nitrogen dioxide.

[8] This means that due to their small size microbes are easily dispersed to all kinds of environments. Then, environmental influences e.g. sun irradiation and temperature decide, if a microorganism thrives or deteriorates [de Wit and Bouvier, 2006].

## 1.3 The primary data source

In 1953 the deoxyribonucleic acid (DNA) has been identified as the structure that carries the genetic information of organisms from all domains of life [Watson and Crick, 1953]. DNA consists of small building blocks called nucleotides: Adenine (A), cytosine (C), guanine (G) and thymine (T). These nucleotides bind together in base pairs (A-T, G-C) and form DNA double strands in form of a helix. The nucleotides are interpreted in triplets and translated into proteins in a process called protein synthesis. The DNA template is first transcribed into ribonucleic acid (RNA) and then translated into proteins. Proteins are the building blocks of cells. In proteomics the structure, function and regulation of the proteins of an organism are studied. The genetic code defines the mapping of DNA triplets. In the process of protein synthesis amino acids are combined to form proteins [Wiltschi and Budisa, 2007]. The standard genetic code is depicted in figure 1.4. There are $4^3 = 64$ combinations that encode $20$ amino acids. Starting in

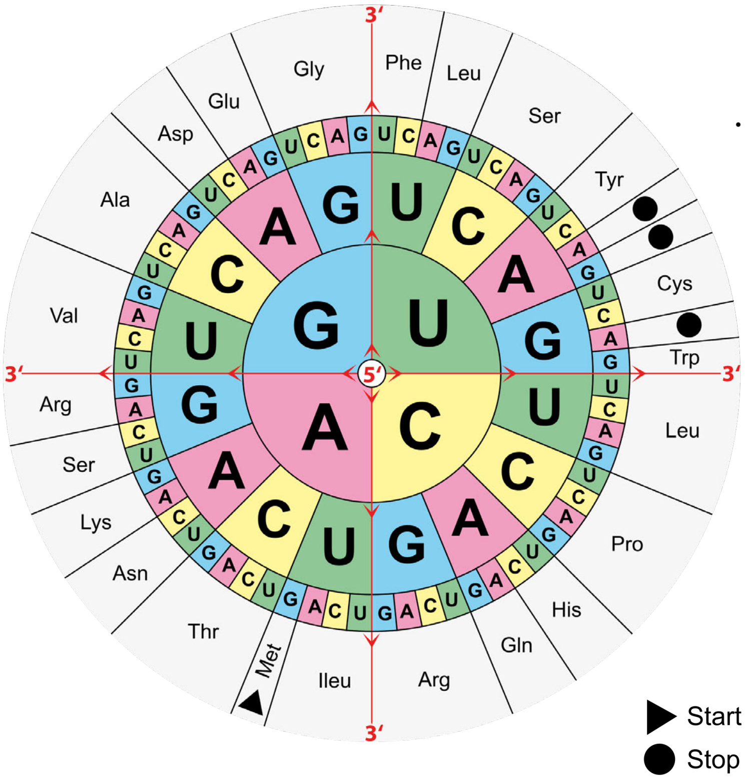


Figure 1.4: The standard genetic code, from (`http://upload.wikimedia.org/wikipedia/commons/7/70/Aminoacids_table.svg`, accessed 30.04.2011) [Wiltschi and Budisa, 2007]

**the middle at the 5' end of the DNA the first nucleotide is read. Then a second, then a third, until the first triplet is complete and the first amino acid has been translated. The process starts over. The triplets are called codons. There is a start codon "AUG" that specifies, where a coding region begins and in which reading frame the DNA has to be interpreted. The DNA is read from 5' to 3' end. A region that can be translated into amino acids is called Open Reading Frame (ORF). There are six potential ORFs, if it is unknown where the 3' end and the 5' are. The start codon assures that all subsequent triplets are correctly interpreted in the right ORF. Furthermore there are three stop codons "UGA", "UAA" and "UAG" that indicate when a coding region is over. The region between a start and a stop codon encodes a gene. When a complete coding region has been translated into amino acids, a gene has been expressed and a protein has been synthesized. The complement of all genes in an organism is called genome [Madigan, 2006]. These genomes contain all biological information[9] that are needed to build and maintain a living organism [Binnewies et al., 2006]. Even though the basic principles how to decode and read genomes have been discovered [Primrose and Twyman, 2003], science is still faraway from comprehensive and thorough understanding of all the functions and processes that are based on these principles. There is a range of disciplines, which study these processes and all the steps[10] involved which are collectively called "omics". Being able to decode and to interpret genomes of organisms allows insight into the 'blueprints' of life, which can be seen as one of the most relevant challenges of our time [Moxon and Higgins, 1997, Henry et al., 2010]. In the field of bioinformatics, which is defined as the use of computational tools to acquire, analyze, store, and access DNA and protein sequences [Madigan, 2006], these sequences are the most important primary data source. Microbial genomics is the discipline which comprises mapping, sequencing, analyzing, and comparing microbial genomes [Madigan, 2006].**

[9] The question arises: "How many genes and proteins does a typical bacterial cell have?" The genome of *Escherichia coli* - a typical bacterium - contains 4.68 million base pairs of DNA. This genome contains roughly 4,300 genes. Some bacterial genomes have three times this number of genes, while some have fewer than one-eighth of this number. A single cell of *Escherichia coli* contains about 1,900 proteins and about 2.4 million total protein molecules [Madigan, 2006].

[10] Included are all steps from the transcription, translation, expression of genes up to the study of metabolism of an organism.

### The ribosome and the 16S and 23S rRNA

Protein synthesis takes place on ribosomes. Ribosomes are cytoplasmic particles composed of ribosomal ribonucleic acid (rRNA) and proteins [Madigan, 2006]. Ribosomes in prokaryotes contain a so-called small subunit RNA (16S[11] rRNA) and large subunit RNA, consisting of the 5S and 23S rRNA [Champney and Kushner, 1976].

To classify (micro)organisms, the 16S and 23S ribosomal ribonucleic acid (RNA) have proven to be useful [Ludwig and Schleifer, 1994, Ludwig et al., 1994]. Due to their important role in protein synthesis, these sequences are under an evolutionary pressure and are thus highly conserved[12]. Consequently, the genes which encode the 16S and 23S rRNA offer a good approximation for phylogeny. The 16S rRNA is more frequently used. This marker gene is furthermore universally present in all domains of life. Marker genes are often stored in curated databases [Pruesse et al., 2007, Ratnasingham and Hebert, 2007] and used to construct phylogenetic trees [Ludwig et al., 2004] that help to classify microorganisms and to speculate about how species might have evolved. Comparison of ribosomal RNA revealed that there are three domans of life: *Bacteria*, *Archaea* and *Eukarya* [Woese et al., 1975, Winker and Woese, 1991] and made it possible for the first time in history to genetically classify microorganisms[13].

### Primary data acquisition - DNA sequencing

Today, since the introduction of the first DNA sequencing methods in 1977 [Sanger et al., 1977, Gilbert and Maxam, 1973], improvements allow the routine sequencing of complete genomes from all domains of life. Obtaining these sequences from microorganisms, however, poses a challenge. In the 20th century - mainly in the medical field - fundamental methods to culture and study microorganisms were developed and applied [Overmann, 2006]. However, 90% - 99% of the microbial diversity have so far resisted isolation attemps [Amann et al., 1995, Curtis et al., 2002]. Consequently, very little is known about microbial diversity yet. Recent approaches like metagenomics and sequencing of

[11] The Svedberg unit "S" is a measure for the sedimentation rate, which is used to differentiate rRNAs.

[12] The 16S and 23S rRNA are yet diverse enough to differentiate organisms based on them.

[13] This was not possible based on their morphology.

marker genes help to overcome this limitation [Handelsman, 2004, Sogin et al., 2006]. In these approaches, bulk DNA is extracted from an environmental sample such as water or sediment and either specific genes are amplified and sequenced or random sequencing is performed. Thus, a fragmented but cultivation-independent overview of an environment's biological diversity and functional potential is provided.

## 1.4 Exponential sequence data growth

Scientists all over the world working in all disciplines of life sciences submit sequence data to the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC is the world's largest public nucleotide sequence data repository (http://www.insdc.org/) and comprised of the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (ENA), and the National Center for Biotechnology Information (NCBI)s GenBank, all of which are daily synchronized.

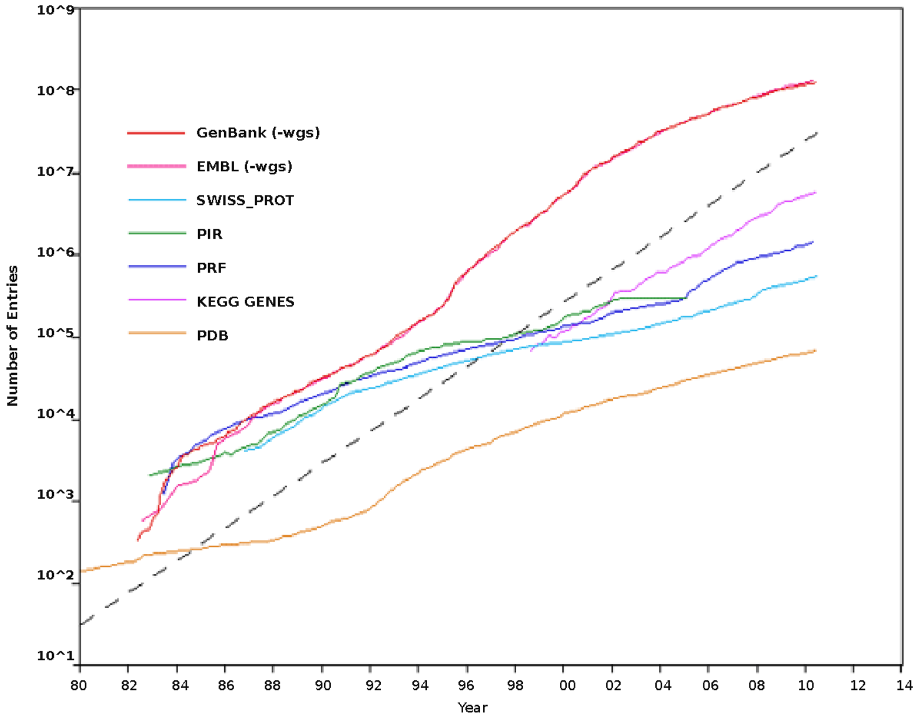


Figure 1.5: DNA sequencing data growth over the past 30 years and into the future. From: `http://www.genome.jp/en/db_growth.html`

Over the last two decades, the amount of sequence data submitted to these public repositories has grown exponentially. Figure 1.5 shows the increase of sequence data in the public data repositories for the INSDC, protein and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. The black dotted line shows that *"the amount of sequence data in the public data repositories has doubled every 18 months"* [Glöckner and Joint, 2010]. Figure 1.6 shows the increase of the DNA sequencing rate.

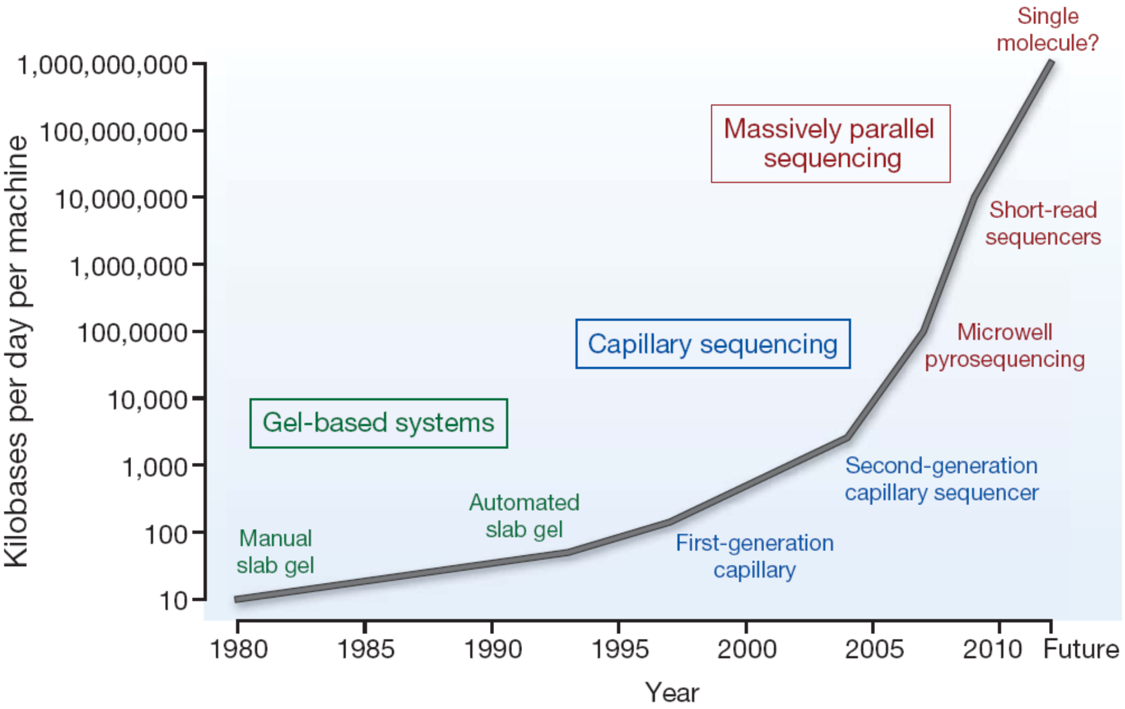


Figure 1.6: Improvements in the rate of DNA sequencing over the past 30 years and into the future. From: [Stratton et al., 2009]

With the advent of Next Generation Sequencing (NGS) [Mardis, 2008], which allows to sequence faster and at lower cost, it can be expected that the sequence accumulation speed will even increase. The increased sequencing rate is already visible in figure 1.6 from 2005 on, which is the year, when the first Next Generation Sequencing (NGS) methods were introduced.

## 1.5 Importance of contextual data

The questions remain: “How can knowledge be derived out of this deluge of data about, what kind of organisms are out there, how many of what kind and what they are doing?" To address these questions, it helps to understand the most important working steps that lead to the generation of sequence data. Figure 1.7 shows a generalized workflow in microbial genomics.

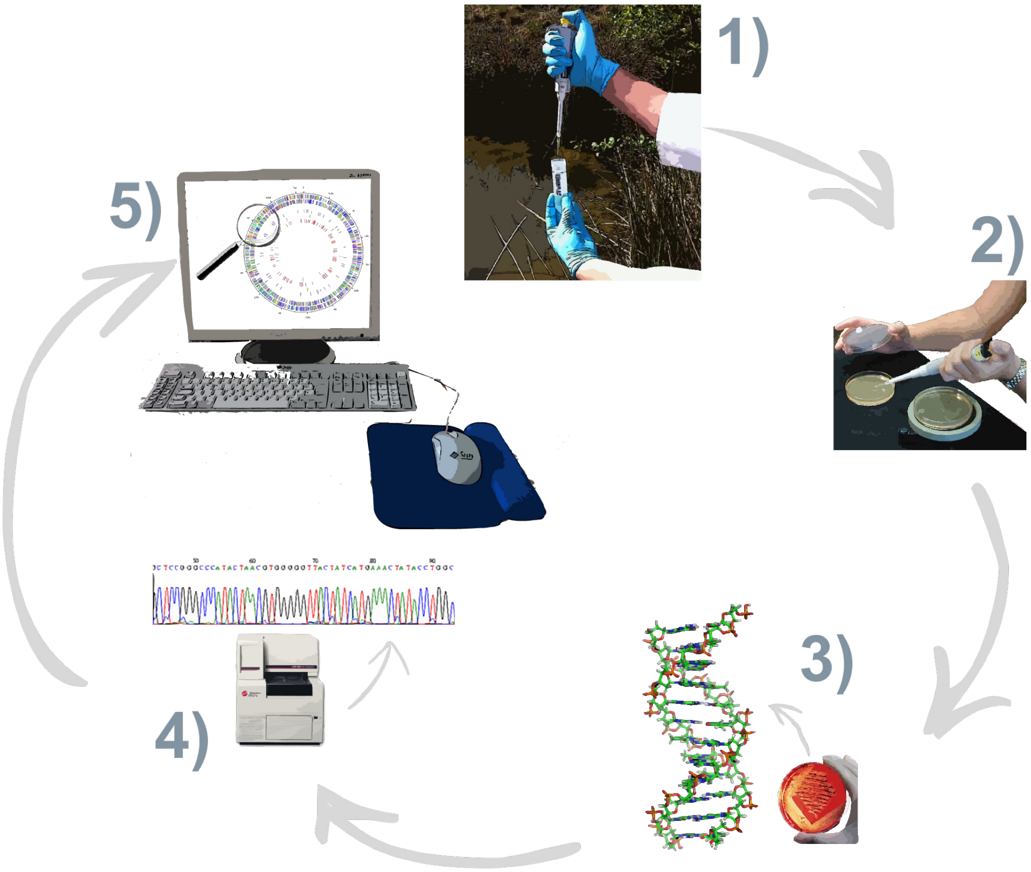


Figure 1.7: A generalized workflow for microbial genomics

When it is aimed to analyze genomes from microorganisms, first, biological samples have to be taken in the field (figure 1.7, step 1). Different isolation strategies are applied depending on the target organism (figure 1.7, step 2). Once in pure culture, DNA is extracted (figure 1.7, step 3) and sequenced (figure 1.7, step 4). Today, complete genomes are often sequenced, assembled and finally analysed (figure 1.7, step 5). The workflow begins very practically in the field, continues with technical work in the laboratory and ends with sequence data analysis on the computer. In metagenomics, DNA is extracted directly from environmental samples, without applying cultivation techniques, and either specific genes are amplified and sequenced or random sequencing is performed. As soon as sequence data is available, it can be submitted to the INSDC.

In all cases, a lot of contextual data is generated along the workflow. It is very important to store contextual data, because they help to understand the context in which the sequence data was created. Often, contextual data are called metadata, which is the more general term. Metadata are recursively defined as data about data or information used to interpret data [Matthew B. Jones, 2006]. Contextual data are data about the environmental context and the processing steps that were applied. These can range from data about the geographic location, sampling time, habitat, or about experimental procedures used to obtain the sequences up to video data recorded during sampling. Especially data about the geographic location (x, y, z) and the point in time (t), when samples are taken from the environment, have proven to be a key data tuple [Martiny et al., 2006].

Normally, these data are recorded in field and laboratory notebooks or in some proprietary file format, such as Microsoft® Excel®. But for others to be able to reconstruct the context, and not only the limited amount of people that have access to these records, it is necessary to make this data publicly available in an electronic form. Some of the data end up in scientific publications, mostly in the form of natural language. To extract these data manually or automatically in text-mining efforts from the scientific literature, is a rather time consuming process [Hirschman et al., 2002]. For example to extract information about the position, where a sample has been taken, publications would have to be automatically screened for keywords like: "position", "location", "GPS", "site", "near", "sampling“, "extraction", "collection“, "specimen" to name a few. The terms listed, demonstrate already the problem with ambiguities in natural language. The term "site" does not necessarily describe a geographic location, but can also be used to describe a DNA binding site. Search patterns like regular expressions that include and exclude certain passages, need to be developed and refined. In many cases the ambiguities will have to be resolved by an expert looking at the search results. This takes time. Given the exponential data accumulation speed, it is valid to assume that efforts in this direction will not be able to keep pace.

Thus, it is of crucial importance to capture and deposit these contextual data along with sequence data in public repositories in an unambiguous and structured way. If these data are missing, key analysis

**possibilities like temporal or spatial comparisons are hampered or simply not possible and comparability of the data cannot be assured. The fact that *“latitude, longitude, and time, elements of the key contextual data tuple (x,y,z,t), are only reported in 7.3% and 7.2% of all submissions“* [Hankeln et al., 2010] shows that the majority of public sequence data is not sufficiently annotated. This fact has been recognized by the Genomic Standards Consortium (GSC), which *”is an open-membership working body which formed in September 2005. The goal of this international community is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data“*(www.gensc.org).**

**The GSC developed a series of checklists to specify which data should be captured and stored along with sequence data [Field et al., 2008]. The INSDC databases support the storage of these parameters. Recently, the life science community has begun to develop tools that implement these standards and to actively integrate these different data sources. *“Data integration is the process of combining data residing at different sources and providing the user with a unified view of these data“* [Lenzerini, 2002].**

**Once contextualized, a far greater scope of analyses can be performed. Studies in various disciplines of life science have already shown the power of contextual data enriched sequence studies. In marine microbiology it could be shown that there are conserved diversity patterns along the depth continuum [DeLong et al., 2006]. Furthermore, annually recurring diversity patterns could be identified in certain regions of the ocean [Fuhrman et al., 2006]. In the medical field the global outbreaks of epidemics can be monitored globally [Janies et al., 2007, Salzberg et al., 2007, Schriml et al., 2010][14]. All these studies exemplify the potential of globally integrated data. The tighter the integration of sequence data with contextual data will be, the easier it will become to carry out sequence data analysis studies in larger contexts. This offers an approach to answer the basic questions ”Who is out there?”, ”How many of which kind?” and ”What are they doing?”. Moreover, knowledge will become obtainable about the complex mechanisms of the Earth’s biosphere on the micro and macro scale.**

[14] There are many more examples that show the increased interpretability of contextualized sequence data: [Tyson et al., 2004, Sogin et al., 2006, Seshadri et al., 2007, Huber et al., 2007, Rusch et al., 2007].

## 1.6 Research aims and motivation

**After the relevance of the field of microbial genomics and the need for contextual and sequence data integration has been identified, the motivation of this thesis was to integrate contextual and sequence data in aid of biological knowledge generation in marine microbial genomics[15]. This has been done by the development of the specialized data integration tools MetaBar and CDinFusion (chapter 2 and 3), by contributing to the development of contextual data standards within the realm of the GSC (chapter 4) and by participating in the development of the marine ecological genomics platform (chapter 5). Finally, a sequence analysis study has been carried out that exemplifies the power of *in silico* knowledge generation in an environmental context (chapter 6).**

## 1.7 Overall thesis structure

**To provide the reader with an overview about the structure of this thesis, the topics of the chapters are sorted into the knowledge pyramid 1.8 introduced in 1.1. The thesis follows a bottom-up approach. It starts on the data level at the base of the pyramid with the consistent contex-**

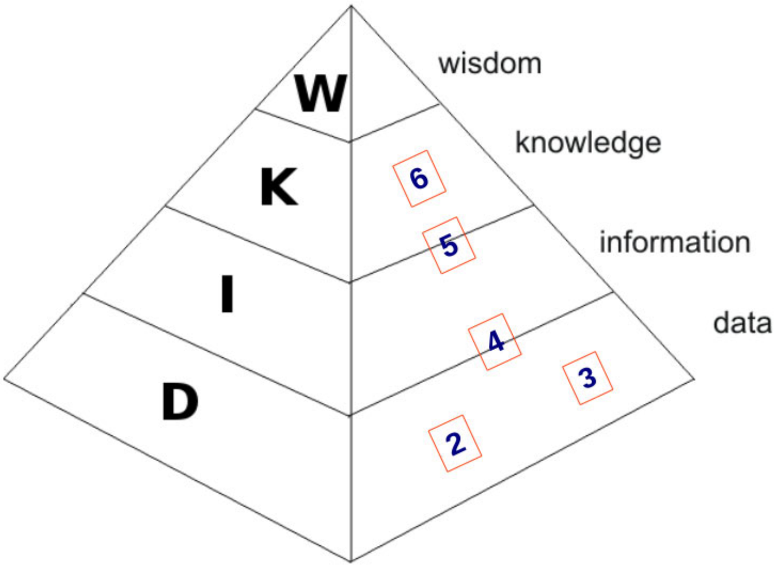


Figure 1.8: Overall thesis structure along the DIKW

[15] This thesis has to be seen as a contribution to this. The realization of the formulated aims will only be possible in a community effort.

tual data acquisition tool MetaBar (chapter 2), which aims to support biologists to capture contextual data from sampling to sequencing. CDinFusion (chapter 3) is a tool to prepare contextualized sequence data submissions. It helps to prepare contextual data enriched sequence submissions to the INSDC. At the interface between data and information is a chapter about the GSC contextual data standard “Minimum Information about a MARKer Sequence” (MIMARKS) (chapter 4), this standard extends and refines the previously published checklist “Minimum Information about a Genome Sequence” (MIGS)/ ”Minimum Information about a Metagenome Sequence“ (MIMS) [Field et al., 2008] that are collectively called MIxS. At the interface between information and knowledge is a chapter about the megx.net platform (chapter 5), an internet portal that provides users with a unified view on integrated sequence, environmental and geographic data. It provides tools to access and analyse these data with graphical user interfaces. The thesis finishes, close to the top of the pyramid, with a chapter about *in silico* hypothesis and knowledge generation, in a study where metagenomic sequence data were analysed in their environmental context (chapter 6).

## 1.8 Publication overview

### 1) MetaBar - A tool for consistent contextual data acquisition and standards compliant submission

Authors: Wolfgang Hankeln, Pier Luigi Buttigieg, Dennis Fink, Renzo Kottmann, Pelin Yilmaz and Frank Oliver Glöckner

Published in: BMC Bioinformatics, June 2010

Personal Contribution: Developed and implemented MetaBar and wrote the initial manuscript.

Relevance: To provide the life science community with a tool that allows to capture contextual data consistently, that accumulate, when samples are collected and processed.

### 2) CDinFusion Submission-ready, on-line Integration of sequence and contextual data

**Authors:** Wolfgang Hankeln, Norma Johanna Wendel, Jan Gerken, Jost Waldmann, Pier Luigi Buttigieg, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner
**Submitted to:** PLoS ONE, April 2011
**Personal Contribution:** Developed and implemented CDinFusion together with Norma Johanna Wendel, Jan Gerken and Jost Waldmann. Wrote the initial manuscript.
**Relevance:** To provide the life science community with a tool to enrich sequence data with contextual data prior to submission to the INSDC.

### 3) The Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

**Authors:** Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison, Philippe Rocca-Serra, Peter Sterk, Manimozhiyan Arumugam, Mark Bailey, Laura Baumgartner, Bruce W Birren, Martin J Blaser, Vivien Bonazzi, Tim Booth, Peer Bork, Frederic D Bushman, Pier Luigi Buttigieg, Patrick S G Chain, Emily Charlson, Elizabeth K Costello, Heather Huot-Creasy, Peter Dawyndt, Todd DeSantis, Noah Fierer, Jed A Fuhrman, Rachel E Gallery, Dirk Gevers, Richard A Gibbs, Inigo San Gil, Antonio Gonzalez, Jeffrey I Gordon, Robert Guralnick, Wolfgang Hankeln, Sarah Highlander, Philip Hugenholtz, Janet Jansson, Andrew L Kau, Scott T Kelley, Jerry Kennedy, Dan Knights, Omry Koren, Justin Kuczynski, Nikos Kyrpides, Robert Larsen, Christian L Lauber, Teresa Legg, Ruth E Ley, Catherine A Lozupone, Wolfgang Ludwig, Donna Lyons, Eamonn Maguire, Barbara A Methé, Folker Meyer, Brian Muegge, Sara Nakielny, Karen E Nelson, Diana Nemergut, Josh D Neufeld, Lindsay K Newbold, Anna E Oliver, Norman R Pace, Giriprakash Palanisamy, Jörg Peplies, Joseph Petrosino, Lita Proctor, Elmar Pruesse, Christian Quast, Jeroen Raes, Sujeevan Ratnasingham, Jacques Ravel, David A Relman, Susanna Assunta-Sansone, Patrick D Schloss, Lynn Schriml, Rohini Sinha, Michelle I Smith, Erica Sodergren, Aymé Spor, Jesse Stombaugh, James M Tiedje, Doyle V Ward, George M Weinstock,

**Doug Wendel, Owen White, Andrew Whiteley, Andreas Wilke, Jennifer R Wortman, Tanya Yatsunenko and Frank Oliver Glöckner**
**Submitted to:** nature biotechnology, accepted April 2011
**Personal Contribution:** Initial talk with the title: "Survey results: MInimal list of contextual data fields for ENvironmental Sequences (MIENS)" at the 6th meeting of the GSC at the EBI (Hinxton, UK) October 2008, which was the starting point for the development of this standard, that was later renamed to MIMARKS. Contributed suggestions for improvements of the data fields, during implementation work of this standard in the tools MetaBar and CDinFusion.
**Relevance:** Standards development for contextual data.

### 4) Megx.net - Integrated database resource for marine ecological genomics

**Authors:** Renzo Kottmann, Ivaylo Kostadinov, Melissa Beth Duhaime, Pier Luigi Buttigieg, Pelin Yilmaz, Wolfgang Hankeln, Frank Oliver Glöckner
**Published in:** Nucleic Acids Research, October 2010
**Personal Contribution:** Involvement in the web portal design and programming, bug fixing and improvement of the environmental data layers in the Genes Mapserver (with Renzo Kottmann, Ivaylo Kostadinov, Pier Luigi Buttigieg, Pelin Yilmaz).
**Relevance:** To enhance the core megx.net concept to the research platform of the Microbial Genomics Group at Max Planck Institute for Marine Microbiology.

### 5) Ecological perspectives on domains of unknown function: a marine point of view

**Authors:** Pier Luigi Buttigieg, Wolfgang Hankeln, Melissa Beth Duhaime, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner
**Submitted to:** ISME Journal, January 2011
**Personal Contribution:** Helped to analyse the data after HMM calculations by generating network graphs together with Pier Luigi Buttigieg.
**Relevance:** Generating hypothesis about the possible functions of protein domains *in silico*, based on their co-occurrence patterns and

**environmental gradients.**

CHAPTER 2

# METABAR

## A tool for consistent contextual data acquisition and standards compliant submission

**Authors:** Wolfgang Hankeln, Pier Luigi Buttigieg, Dennis Fink, Renzo Kottmann, Pelin Yilmaz and Frank Oliver Glöckner

**Published in:** BMC Bioinformatics, June 2010

**Personal Contribution:** Developed and implemented MetaBar and wrote the initial manuscript.

**Relevance:** To provide the life science community with a tool that allows to capture contextual data consistently, that accumulate, when samples are collected and processed.

## 2.1 Abstract

**Background:** Environmental sequence datasets are increasing at an exponential rate; however, the vast majority of them lack appropriate descriptors like sampling location, time and depth/altitude: generally referred to as metadata or contextual data. The consistent capture and structured submission of these data is crucial for integrated data analysis and ecosystems modeling. The application MetaBar has been developed, to support consistent contextual data acquisition.

**Results:** MetaBar is a spreadsheet and web-based software tool designed to assist users in the consistent acquisition, electronic storage, and submission of contextual data associated to their samples. A preconfigured Microsoft® Excel® spreadsheet is used to initiate structured contextual data storage in the field or laboratory. Each sample

**is given a unique identifier and at any stage the sheets can be uploaded to the MetaBar database server. To label samples, identifiers can be printed as barcodes. An intuitive web interface provides quick access to the contextual data in the MetaBar database as well as user and project management capabilities. Export functions facilitate contextual and sequence data submission to the International Nucleotide Sequence Database Collaboration (INSDC), comprising of the DNA DataBase of Japan (DDBJ), the European Molecular Biology Laboratory database (EMBL) and GenBank. MetaBar requests and stores contextual data in compliance to the Genomic Standards Consortium specifications. The MetaBar open source code base for local installation is available under the GNU General Public License version 3 (GNU GPL3).**

**Conclusion: The MetaBar software supports the typical workflow from data acquisition and field-sampling to contextual data enriched sequence submission to an INSDC database. The integration with the megx.net marine Ecological Genomics database and portal facilitates georeferenced data integration and metadata-based comparisons of sampling sites as well as interactive data visualization. The ample export functionalities and the INSDC submission support enable exchange of data across disciplines and safeguarding contextual data.**

## 2.2 Background

The technological advancement in molecular biology facilitates investigations of biodiversity and functions on a temporal and geospatial scale. Improved sampling and laboratory methods, together with fast and affordable sequencing technologies [Hall, 2007], provide the framework to create a network of data points capable to answer basic ecological questions such as: ‘Who is out there?’ and ‘What are these organisms doing?’ To shed light on the complex interplay, adaptation and survival mechanisms of organisms in times of global change, contextual data describing the surrounding environment of sampling locations are of crucial importance [Field et al., 2008]. At the very least, the latitude and longitude (x, y),the depth/altitude (z) in relation to sea level, and the sampling date and time (t) must be provided to allow anchoring

**molecular sequence data to their environmental context. If every sequence entry in the INSDC databases, comprising of DDBJ, EMBL and GenBank, would be thus georeferenced, researchers would have the post factum opportunity to contextualize these sequences with environmental data [Wieczorek et al., 2004]. The power of contextual data enriched sequence data sets for the environmental and medical field has been recently documented [DeLong et al., 2006, Janies et al., 2007, Ramette, 2007, Fuhrman et al., 2006, Pommier et al., 2007, Parks et al., 2009, Green et al., 2008, Schriml et al., 2010, Field, 2008].**

**Unfortunately, a survey in the EMBL sequence repository has shown that only a minor set of sequences are accompanied by a relevant amount of contextual data. For example, latitude, longitude (INSDC: lat_lon), and time (INSDC: collection_date), elements of the key contextual data tuple (x,y,z,t), are only reported in 7.3% and 7.2% of all submissions [Guy Cochrane, personal communication, October 2009]. But even if these data are available, correctness is not guaranteed.**

**The paucity of sequence associated contextual data has been recognized by the primary database providers and biocuration efforts are currently underway for specific subsets. The National Center for Biotechnology Information (NCBI), for example, curates the Reference Sequence (RefSeq) database which aims to provide a comprehensive, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts and proteins (`http://www.ncbi.nlm.nih.gov/RefSeq/`). The European Molecular Biology Laboratory (EMBL) provides the UniProt/Swiss-Prot Knowledgebase which focuses on high quality protein sequence annotations (`http://www.ebi.ac.uk/uniprot/`) [Consortium, 2010]. However, the common aim of these efforts is to enhance the quality of the sequence or protein data and annotations rather than to provide more information on the data processing or the environment where the sample or organism has been taken.**

**To improve the quantity and quality of contextual data describing the environment of a sample is currently addressed by several projects which systematically collect georeferenced sequence data, environmental parameters, and further curated metadata [Howe et al., 2008]. SILVA [Pruesse et al., 2007] or RDP II [Cole et al., 2005] are examples for specialized databases that offer users curated and quality checked ribosomal RNA sequences that are often enriched with more reliable**

**contextual and taxonomic information than originally annotated by the sequence submitters. Furthermore, there are projects which curate the contextual data associated to the primary sequence data to facilitate specific analysis purposes. For example the Genomes OnLine Database (GOLD) collects metadata for ongoing and completed genome sequencing projects [Liolios et al., 2008]. The Visualization and Analysis of Microbial Population Structures (VAMPS) project, with its integrated collection of tools for researchers, aims to visualize and analyze data for microbial population structures and distributions. All the contextual data in VAMPS comes from the MICROBIS database management system of the International Census of Marine Microbes (ICoMM: `http://icomm.mbl.edu/microbis/`). The megx.net portal (`www.megx.net`) [Kottmann et al., 2010] systematically integrates environmental parameters and sequence data of marine microbial genomes and metagenomes using georeferencing as an anchor. In 2005 the international Genomic Standards Consortium (GSC) introduced checklists to promote standardized contextual data acquisition and storage. So far the Minimum Information about a Genome (Metagenome) Sequence (MIGS/MIMS) has been published [Field et al., 2008] and the Minimum Information about an Environmental Sequence (MIENS) is in development (`http://gensc.org/gc_wiki/index.php/MIENS`). For data exchange, the Genomic Contextual Data Markup Language (GCDML) [Kottmann et al., 2008] has been developed. A corollary of these ongoing efforts is the need to support field scientists in the consistent capture, storage and submission of both contextual and sequence data. Handlebar, a lightweight Laboratory Information Management System (LIMS) for the management of barcoded samples, in part addresses this issue by supporting the acquisition and processing of contextual data compliant with GSC standards [Booth et al., 2007]. The Barcode of Life Database (BOLD) initiative, which aims to identify and classify all eukaryotic life on Earth [Ratnasingham and Hebert, 2007], also includes an advanced data acquisition and submission system. Unfortunately the system only supports phylogenetic markers which serve the eukaryotic domain e.g. the cytochrome c oxidase I (COI), which is only present in Eukarya and absent in the other domains of life, and so far exclude Archaea and Bacteria. Furthermore, it does not support the**

**printing of database identifiers as barcodes to label collected samples. Even though initiatives and tools exist to enhance the quantity and quality of contextual data subsequent to sequence submission, the amount of contextual data in the INSDC databases remains an issue. In summary, the most likely reasons for the persisting scarcity of consistent contextual data are: (1) Contextual data that are recorded in the field are often not stored electronically in structured databases. Consequently contextual data get rapidly unlinked from the sequence data and finally 'forgotten' in the sequence submission process. (2) There is a lack of automatic quality checking mechanisms active before data submission. Unfortunately, the flood of data entering the public databases prevents any manual curation process. (3) The sheer amount of potential contextual data with respect to the different fields of research ranging from textual data to images or even videos would rapidly exceed the capacities of the INSDC databases. Consequently, only a commonly agreed and standardized subset of data can be stored and made available.**

**Here the user-centric, web-based tool MetaBar is presented. MetaBar offers all the required features for sample identification and barcode labeling allowing robust sample tracking and inventorying. MetaBar is focused on the acquisition of contextual data recorded during sampling in the field 'offline' using spreadsheets. All recorded contextual data can be subsequently uploaded and consistently stored in an underlying database. The web Graphical User Interface (GUI) provides advanced user management and access to data and barcodes. Vitally, the tool captures GSC standards compliant data and it is integrated into a set of tools to facilitate further data usage such as integration, visualization and analysis available from the Marine Ecological Genomics database and portal, megx.net. Finally, MetaBar supports contextual data enriched sequence submission to the INSDC databases. The tool is not restricted to any given research field or domain of life, but can universally be applied to capture the contextual data of any biological sample.**

**It is designed to support the complete workflow from the sampling event up to the sequence submission to an INSDC database.**

## 2.3 Implementation

### Programming languages, tools and frameworks

MetaBar is programmed in the object-oriented, platform-independent programming language, Java 1.5
(http://www.java.com/en/). MetaBar is a multiuser web application using Apache Tomcat (http://tomcat.apache.org/), the open source Spring framework (http://www.springsource.org/about), jasig CAS (http://www.jasig.org/cas), which is used as a central authentication service to implement the user management, and Apache POI (http://poi.apache.org/) to parse the Microsoft® Excel® spreadsheets used for data input.
Any Java objects generated are stored in a PostgreSQL database using the iBATIS persistence framework
(http://ibatis.apache.org/). The input fields in the Microsoft® Excel® spreadsheet are validated using Visual Basic (VBA) macros. The web interface has been continuously tested during development using Selenium IDE (http://seleniumhq.org/projects/ide/) and JUnit (http://www.junit.org/). The source code of the MetaBar application is available under the GNU GPL3 (http://www.megx.net/metabar/data/metabar-1.0.tar.gz) and as additional file 1 to this publication.

### Core software components

The MetaBar application consists of (1) the Microsoft® Excel® acquisition spreadsheet which is used to capture and auto-correct the contextual data, (2) the MetaBar server which generates and receives the acquisition spreadsheets, parses the data and stores them in (3) MegDB, a PostgreSQL database which is the central database of the megx.net portal [Kottmann et al., 2010].

### External software components

MetaBar is integrated into a set of external tools directly accessible from the web interface. The interpolation of environmental physical and chemical parameters of the oceans can be initiated via the WOA05 data extractor of the megx.net portal. On the fly visualization of sampling sites on a world map can be performed using the Genes Mapserver (http://www.megx.net/gms) and in Google Earth® via the

KML export function. The data can be exported prior to sequence submission as a structured comment block for submission to INSDC by the Sequin tool (`www.ncbi.nlm.nih.gov/Sequin/index.html`). MetaBar also includes a data export to GCDML [Kottmann et al., 2008] for report creation and data exchange.

## 2.4 Results

The core application can be best explained by describing the workflow across the different MetaBar components (Figure 2.1). First, users log on to the MetaBar web server (Figure 2.1, step 1). Upon entry, users can allocate a certain range of sample identifiers before, during or after a sampling campaign. The identifier consists of a six digit [sample-id] that is incremented with every new sample, a six digit [project-id] that is incremented with every new project and a two digit [institute-

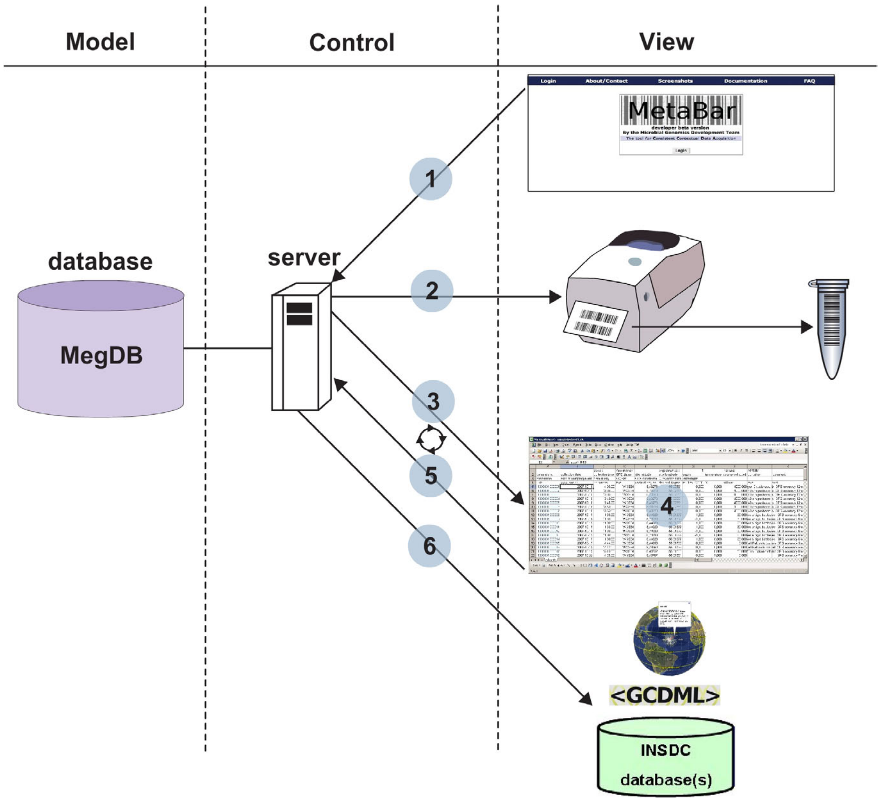
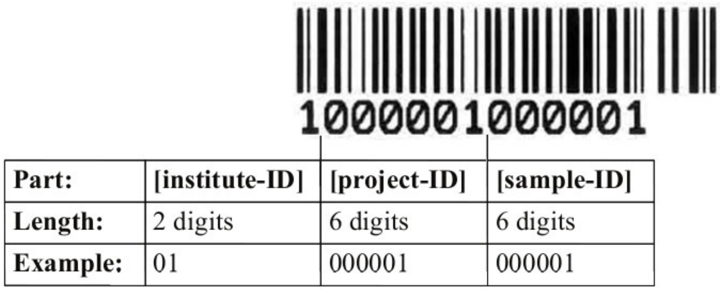


Figure 2.1: Scheme of the MetaBar workflow. Users are allowed to create barcodes, print them onto physical labels, and to capture, upload, and update contextual data for the barcodes using Excel® spreadsheets. The contextual data can be exported to various formats and submitted to the INSDC databases.

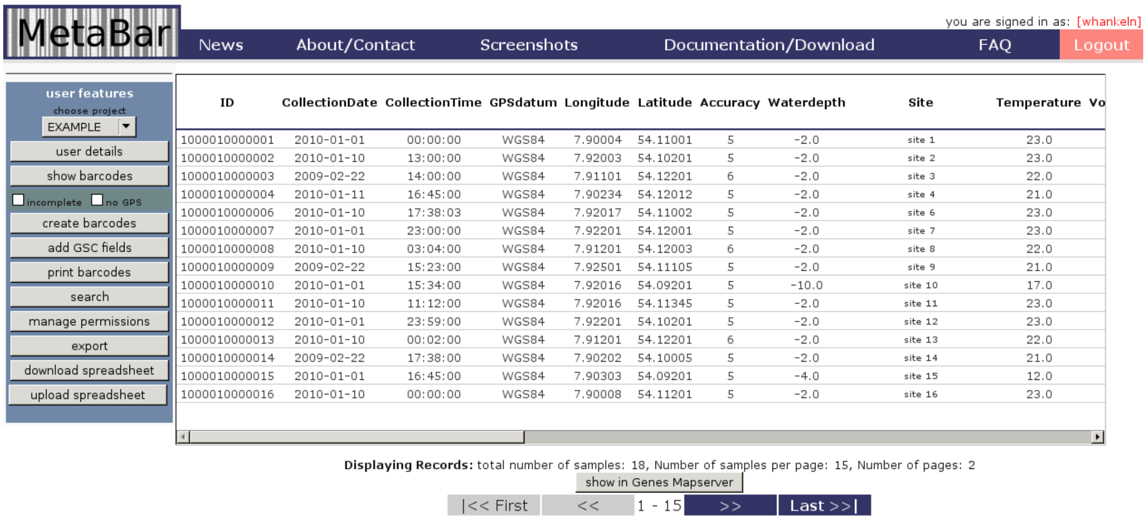
**id] that is fixed and identifies a certain institute. The combination of these three parts in one identifier assures the unique identification of each sample. The identifiers can be printed (Figure 2.1, step 2) as barcodes onto labels (Figure 2.2) that can be placed on sample containers and pasted into laboratory notebooks for consistency. Users can download (Figure 2.1, step 3) the acquisition spreadsheet containing the allocated identifiers and the empty contextual data fields in the first worksheet. As the user fills these fields, VBA validation macros check the inputs and users are prompted to use, for example, correct formats in the correct numerical range, where applicable. New worksheets can be added to the spreadsheet. Thus, any additional data outside of the MetaBar model can be added to the same file. Once the worksheets are filled (Figure 2.1, step 4) they can be uploaded (Figure 2.1, step 5) to the MetaBar web server. After the upload is finished, the file is parsed and the values in the first worksheet are stored in the respective relational fields of the central database. The additional worksheets in the file are not lost, but stored as binary data in the database. The latter three steps can be repeated whenever it is necessary to edit and update the data. Users can log in to the system at any time to search and browse their data via the web GUI (Figure 2.3). Additionally, the MetaBar core set of contextual data fields can be extended for each sample with further GSC compliant parameters. These additional fields are organized into different types of report and environmental packages, each containing further parameters.**



| Part: | [institute-ID] | [project-ID] | [sample-ID] |
|---|---|---|---|
| Length: | 2 digits | 6 digits | 6 digits |
| Example: | 01 | 000001 | 000001 |

Figure 2.2: The barcode identifier. Barcodes consist of three parts ([institute-ID], [project-ID] and [sample-ID]) which in combination uniquely identify a sample. The barcodes are printed onto physical labels that can be placed on sample containers.

**The parameters can be directly selected and updated via the web interface (Figure 2.4). MetaBar can also be used as an inventory e.g. for freezer contents. The database may be queried using sample identifiers**



| ID | CollectionDate | CollectionTime | GPSdatum | Longitude | Latitude | Accuracy | Waterdepth | Site | Temperature | Vo |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000010000001 | 2010-01-01 | 00:00:00 | WGS84 | 7.90004 | 54.11001 | 5 | -2.0 | site 1 | 23.0 | |
| 1000010000002 | 2010-01-10 | 13:00:00 | WGS84 | 7.92003 | 54.10201 | 5 | -2.0 | site 2 | 23.0 | |
| 1000010000003 | 2009-02-22 | 14:00:00 | WGS84 | 7.91101 | 54.12201 | 6 | -2.0 | site 3 | 22.0 | |
| 1000010000004 | 2010-01-11 | 16:45:00 | WGS84 | 7.90234 | 54.12012 | 5 | -2.0 | site 4 | 21.0 | |
| 1000010000006 | 2010-01-10 | 17:38:03 | WGS84 | 7.92017 | 54.11002 | 5 | -2.0 | site 6 | 23.0 | |
| 1000010000007 | 2010-01-01 | 23:00:00 | WGS84 | 7.92201 | 54.12001 | 5 | -2.0 | site 7 | 23.0 | |
| 1000010000008 | 2010-01-10 | 03:04:00 | WGS84 | 7.91201 | 54.12003 | 6 | -2.0 | site 8 | 22.0 | |
| 1000010000009 | 2009-02-22 | 15:23:00 | WGS84 | 7.92501 | 54.11105 | 5 | -2.0 | site 9 | 21.0 | |
| 1000010000010 | 2010-01-01 | 15:34:00 | WGS84 | 7.92016 | 54.09201 | 5 | -10.0 | site 10 | 17.0 | |
| 1000010000011 | 2010-01-10 | 11:12:00 | WGS84 | 7.92016 | 54.11345 | 5 | -2.0 | site 11 | 23.0 | |
| 1000010000012 | 2010-01-01 | 23:59:00 | WGS84 | 7.92201 | 54.10201 | 5 | -2.0 | site 12 | 23.0 | |
| 1000010000013 | 2010-01-10 | 00:02:00 | WGS84 | 7.91201 | 54.12201 | 6 | -2.0 | site 13 | 22.0 | |
| 1000010000014 | 2009-02-22 | 17:38:00 | WGS84 | 7.90202 | 54.10005 | 5 | -2.0 | site 14 | 21.0 | |
| 1000010000015 | 2010-01-01 | 16:45:00 | WGS84 | 7.90303 | 54.09201 | 5 | -4.0 | site 15 | 12.0 | |
| 1000010000016 | 2010-01-10 | 00:00:00 | WGS84 | 7.90008 | 54.11201 | 5 | -2.0 | site 16 | 23.0 | |

**Displaying Records:** total number of samples: 18, Number of samples per page: 15, Number of pages: 2

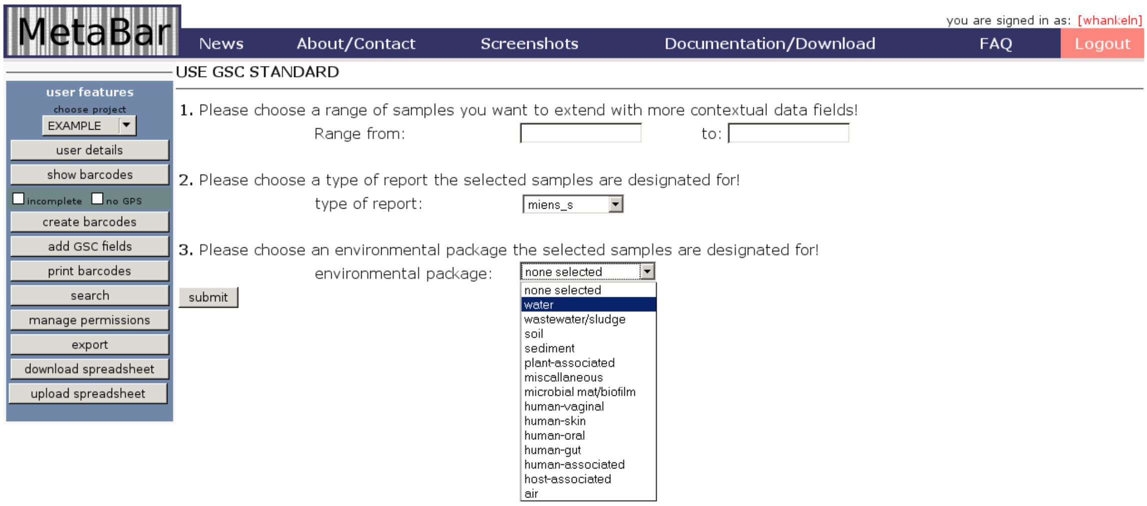
Figure 2.3: Screenshot of the graphical user interface I. Uploaded contextual data can be browsed and queried online.

by scanning their barcodes with an appropriate device, by manually entering their corresponding numeric code, or by text search on a metadata field. The query then retrieves all corresponding contextual data stored in the system. MetaBar is integrated into a set of external tools with direct access from the web interface.

The interpolation of physical and chemical parameters such as temperature, nitrate, phosphate, salinity, silicate, dissolved oxygen, oxygen saturation, apparent oxygen utilization and chlorophyll of a marine sampling site can be initiated via the WOA05 data extractor of the megx.net portal. On the fly visualization of sampling sites on a world map can be performed using the Genes Mapserver (`http://www.megx.net/gms`). Furthermore, four export functions (Figure 2.1, step 6) are currently supported: (1) an export to KML to visualize sampling sites including their contextual data in Google Earth®, (2) an export to GCDML [Kottmann et al., 2008] for report creation and for data exchange, (3) an export to a GSC compliant MIGS/MIMS/MIENS spreadsheet, and (4) an export as a structured comment for sequence data submission to the INSDC databases using the Sequin tool (`www.ncbi.nlm.nih.gov/Sequin/index.html`).

### Role, ownership and permission concept

MetaBar's user management provides a "MetaBar admin", a "project admin" and a "MetaBar user" mode. These modes depend on the role



USE GSC STANDARD

1. Please choose a range of samples you want to extend with more contextual data fields!
Range from: to:

2. Please choose a type of report the selected samples are designated for!
type of report: miens_s

3. Please choose an environmental package the selected samples are designated for!
environmental package: none selected

Figure 2.4: Screenshot of the graphical user interface II. Contextual data entries can be extended with GSC parameters. All contextual data can be exported and submitted to the INSDC databases.

that is assigned to a certain user account. The web GUI possesses a cascading menu on the left which contains the “MetaBar admin features”, the “project admin features”, and the “MetaBar user features”, respectively. Furthermore, a sophisticated ownership and permission concept offers the users to share their data with other users in the same project giving them read or write permissions, or to prevent access for others. Project admins have the possibility to transfer the ownership of a set of samples to another user, create projects, assign users to a project and to remove users from a project. For a given MetaBar installation there is only one MetaBar admin who can create users, assign or dismiss project admins and delete samples or whole projects. A quick reference guide describing the general workflow of MetaBar from the acquisition to the submission of data is available on the website. Examples for metadata enriched INSDC database entries, created with MetaBar, are available through the accession numbers: [GenBank:GU949561 and GenBank:GU949562].

## 2.5 Discussion

MetaBar can be used whenever it is necessary to capture contextual data that describe the environmental origin of a sample. The system has been tested in several studies in close collaboration with biologists taking samples in the field. By integrating their feedback MetaBar should qualify as user-friendly, scientist-centric software tool.

### Case study

The above mentioned studies allow the typical MetaBar workflow to be generalized as follows. A scientist acting as the project administrator (PA) plans a sampling campaign together with two members in his research team (PM1, PM2). The project EXAMPLE is created in MetaBar and the users PM1 and PM2 are added to EXAMPLE. It is anticipated that PM1 will collect five push core samples of sediment while PM2 will collect five water samples. Thus, PM1 and PM2 create five barcodes each and download their acquisition spreadsheets. These barcodes are printed in multiple copies to label sample containers, appear in field notebooks and for contingency.

During sample collection, PM1 and PM2 log contextual data such as latitude, longitude, depth and sampling time next to the barcode labels pasted in their field notebooks. The sample containers are labeled with the corresponding barcodes and transported back to the laboratory. The link between sample and environment is thus established. At the end of the sampling campaign the contextual data gathered in the field are transferred to PM1 and PM2's acquisition spreadsheets. Barcodes pasted on the sample, the field records and present in the spreadsheet ensure fidelity and the data are then uploaded to the MetaBar server over the internet. PM1 and PM2 may enter further contextual data specific to their sampling environments by selecting the relevant GSC-compliant metadata packages (e.g. “sediment” and “water”, respectively) through the web GUI. The PA and both members of the project can now review the consolidated contextual data for errors or missing values. Corrective action at this stage improves the quality of the data prior to submission.

During laboratory processing, every new subsample is labeled with a copy of the original sample's barcode, preserving the link to the in situ sampling event. Native laboratory protocols and practices are otherwise unaffected and are documented in laboratory books. PM1 sequences the genomes of several sediment sample isolates and PM2 sequences microbial metagenomes from the community in the water sample. Congruent to the environmental extensions, GSC packages corresponding to various study types are available. PM1 and PM2 may use the “MIENS culture (miens_c)” and “metagenome (me)” packages, respectively, to record data specific to their study type (Figure 4). PM1 and PM2 receive their genome and metagenome sequences as FASTA files with automatically generated sequence identifiers in the header. The researchers enter these identifiers into the “seqID” field in the acquisition spreadsheet and export the data to a format for submission to INSDC. With this mapping, these contextual datasets can easily be combined with one or more FASTA sequences using a suitable submission tool. The researchers then submit their metadata-enriched sequences from the EXAMPLE project to an INSDC database. MetaBar implements a neat trade-off between universality and specificity. The export functions assure that the collected data can be publicly stored and shared with the scientific community.

### Comparison of MetaBar and Handlebar

The idea of uniquely identifying samples and storing data about these samples in databases is not new and is widely used in many applications and disciplines. However, tools able to capture the contextual data of environmental samples combined with barcode labeling are rare. To our knowledge with the exception of MetaBar, the only open source tool using barcoding to identify georeferenced samples from the environment is Handlebar [Booth et al., 2007]. A tabular comparison of the programs’ general features can be found in Table 2.1.

HandleBar, as a lightweight LIMS, not only covers contextual data that are recorded during sampling, but also aims to document subsequent sample processing steps in the laboratory. In this respect, MetaBar is a simplification focusing only on the capture of contextual data in the field. MetaBar does not seek to replace well established laboratory bookkeeping or professional LIM systems, but rather aims to complement this process to ensure that contextual data are electronically accessible. Nevertheless, users may choose to use the tool as a storage inventory manager or to store intermediate results of sample processing because it is possible to store additional data in the spreadsheets. It is important to note that coupling contextual data with sequence data before submission to the INSDC databases is a unique feature of MetaBar.

The barcodes are, by concept, solely used to link environmental sam-

| | **Handlebar** | **MetaBar** |
|---|---|---|
| Focus | Web-based lightweight LIMS for handling barcoded samples | Web-based tool for consistent contextual data acquisition with barcoded samples |
| System requirements | Operating system: Windows® or GNU Linux Apache, Perl, PostgreSQL, OpenOffice or Microsoft® Excel® | Without local MetaBar server installation: Operating system: Windows® Internet connection, web browser (e.g. Firefox), Microsoft® Excel® 2003 or higher Optional: EPL barcode printer (e.g. a Zebra® TLP 2824) With local MetaBar server installation: Operating system: Windows® or GNU Linux Apache, Java, Spring, jasig CAS, PostgreSQL, Microsoft® Excel® 2003 or higher |
| Coverage | Metadata that emerges during sampling events and subsequent processing step data | Contextual data that emerges during sampling events (other data optional) |
| Sample type templates | Various | One generic and extensible template |
| Input validation | Done by the server | Done by VBA® macros in the acquisition spreadsheet and on the server |
| Integration into data analysis tool set | GenQuery | `http://www.megx.net` |
| Export functions | - | GCDML, KML (for Google Earth) |
| Contextual data enriched sequence submission support | - | Export to MIGS/MIMS/MIENS and structured comment |

Table 2.1: Features of Handlebar and MetaBar

**ples to contextual and, if available, sequence and species data derived from a labeled sample, thus, no hierarchy or processing method is encoded in the identifiers. Also, sample hierarchies and complex identifier schemes are avoided. This concept does not interfere with native laboratory sample tracking methods, yet ensures consistency in environmental contextual data capture. It is important that users have the flexibility to cover different sample types. MetaBar offers a single template in which a restricted part is parsed to the database and an unrestricted part of the spreadsheet can be changed to contain sample specific additional data. HandleBar offers a set of non-constrained sample templates depending on the sample type and also individual templates can be created. In MetaBar each sample can be extended with further parameters organized into types of report and environmental packages suggested by the GSC.**

**In contrast to HandleBar, data entered into MetaBar's acquisition spreadsheet is validated on input, ensuring correct format before upload to the MetaBar server. This avoids frequent rejection of the acquisition sheet. In HandleBar the validation is done by the web server and erroneous sheets have to be corrected retrospectively by the uploading user.**

**The variety of export features are currently unique to MetaBar.**

**MetaBar is integrated into the megx.net tool set and connected to MegDB. This offers opportunity to work with the data and to analyze them alone, or in the context of other research project data stored in the megx.net database. This level of integration necessitated a user authentication and authorization management system and SSL encryption. Consequently, the local installation of MetaBar requires modification of the open source code base. The software and a detailed installation manual are available at** `www.megx.net/metabar`**. However, accounts on the MetaBar installation hosted at the MPI for Marine Microbiology in Bremen can easily be given to interested users and an "anonymous" project exists where data of external users can be stored anonymously. It is the intention of the Microbial Genomics and Bioinformatics Group at the MPI-Bremen to support this tool as open source in the future.**

### Applicability

**MetaBar has been developed at the Max Planck Institute for Ma-**

rine Microbiology; however, the tool may be readily applied to a wide range of research fields outside the marine sciences. Contextual data fields relevant to air, host associated, human associated, sediment, soil, wastewater sludge or water samples are available via the "add GSC fields" function. The parameters in each of these environmental packages have been selected based on community usage and consensus (`http://gensc.org/gc_wiki/index.php/MIENS`).
For example, fields requesting data on barometric pressure, carbon dioxide, carbon monoxide, chemical administration, humidity, methane, organism count, oxygen, oxygenation status of sample, perturbation, pollutants, respirable particulate matter, sample salinity, sample storage duration, sample storage location, sample storage temperature, solar irradiance, temperature, ventilation rate, ventilation type, volatile organic compounds, wind direction, and wind speed would be presented to users using the air environmental package. Users may easily add new, custom fields as columns using standard Microsoft® Excel® operations. Combined, the GSC extensions and freedom for customization generalize MetaBar's applicability to any scenario necessitating the capture of contextual data describing a sample's environmental origin.

## 2.6 Conclusion

MetaBar offers an integrated contextual data acquisition, storage, and submission solution to the INSDC system. The impact of better contextual data availability and correctness in the primary sequence databases will greatly improve the possibilities to reach a higher level of data integration and interpretation to address basic ecological questions.
MetaBar's integration into the megx.net tool set and its export mechanisms offer extended analysis possibilities via comparison to other scientific studies and with complementary interpolated environmental data. The visualization of the sampling sites on the Genes Mapserver and in Google Earth® offers the users a simple way to show sampling events on the globe and to relate them to other publicly available scientific studies.

Statistical analysis of phylogenetic and functional biodiversity in their environmental context will reveal new insights into the biogeography and habitat adaptation of organisms.
In the medical field, for example, it will be possible to create detailed disease maps which reveal mutation patterns of a certain pathogenic organism over time [Janies et al., 2007,Schriml et al., 2010]. Such maps might help to predict the dispersal of epidemics and pandemics around the globe. For marine microbiology, Ed DeLong and coworkers have successfully shown that there is a stratification of genomic variability along the depth continuum in the water column at a specific sampling location [DeLong et al., 2006]. It has also been demonstrated that specific diversity patterns are annually recurring [Fuhrman et al., 2006]. A dense network of data points, enriched with contextual data, will lead to new insights into the complex interplay of organisms by comparing different sampling sites around the globe and over time. The denser this network of data points, the more will be revealed about the influence of the biotic factor in the elementary nutrient cycles that profoundly affect Earth's climate.

## Availability and Requirements

### Software

Project name: MetaBar

Project homepage: `http://www.megx.net/metabar`

Operating systems: Linux and Windows

Programming language: Java JRE 1.5 or higher

Other requirements: Microsoft® Excel® 2003 or higher, Google Earth® (optional)

License: GNU General Public License version 3 (GNU GPL3)

### Hardware

At least 1024 Mb of RAM

EPL barcode printer (e.g. a Zebra TLP 2824) (optional)

Barcode handscanner (optional)

The software can be tested anonymously using the login: "anonymous" with the password: "testmetabar".

## Authors' contribution

WH developed and implemented MetaBar and wrote the manuscript.

**RK advised programming design and helped with the integration of MetaBar with MegDB and megx.net. DF tested the software on cruises and provided feedback for design improvements. PY assured the MIGS/MIMS/MIENS standard compliance in MetaBar. PLB critically revised the manuscript and took care of EnvO integration. PY, PLB and RK tested the tool in the field. FOG supervised the work and helped with writing the manuscript. All authors read and approved the final manuscript.**

## Acknowledgments

**Thanks to Jens Harder, Christine Klockow, Mirja Meiners and all the testers. Thanks to Dawn Field and Tim Booth at NERC, UK for feedback and useful input during the design phase of MetaBar. Thanks to Norma Wendel for her help in finalizing the MetaBar source code. The study was supported by the Max Planck Society.**

CHAPTER 3

# CDINFUSION

## Submission-ready, on-line Integration of sequence and contextual data

**Authors:** Wolfgang Hankeln, Norma Johanna Wendel, Jan Gerken, Jost Waldmann, Pier Luigi Buttigieg, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner

**Submitted to:** PLoS ONE, April 2011

**Personal Contribution:** Developed and implemented CDinFusion together with Norma Johanna Wendel, Jan Gerken and Jost Waldmann. Wrote the initial manuscript.

**Relevance:** To provide the life science community with a tool to enrich sequence data with contextual data prior to submission to the INSDC.

## 3.1 Abstract

State of the art (DNA) sequencing methods applied in "Omics" studies grant insight into the 'blueprints' of organisms from all domains of life. Sequencing is carried out around the globe and the data is submitted to the public repositories of the International Nucleotide Sequence Database Collaboration. However, the context in which these studies are conducted often gets lost, because experimental data, as well as information about the environment are rarely submitted along with the sequence data. If these contextual or metadata are missing, key opportunities of comparison and analysis across studies and habitats are hampered or even impossible. To address this problem, the

Genomic Standards Consortium (GSC) promotes checklists and standards to better describe our sequence data collection and to promote the capturing, exchange and integration of sequence data with contextual data. In a recent community effort the GSC has developed a series of recommendations for contextual data that should be submitted along with sequence data.

To support the scientific community to significantly enhance the quality and quantity of contextual data in the public sequence data repositories, specialized software tools are needed. In this work we present CDinFusion, a web-based tool to integrate contextual and sequence data in (Multi)FASTA format prior to submission. The tool is open source and available under the Lesser GNU Public License 3. A public installation is hosted and maintained at the Max Planck Institute for Marine Microbiology at `http://www.megx.net/cdinfusion`. The tool may also be installed locally using the open source code available at `http://code.google.com/p/cdinfusion`.

## 3.2 Introduction

The introduction of the first deoxyribonucleic acid (DNA) sequencing methods in 1977 marked a major breakthrough in life science [Gilbert and Maxam, 1973,Sanger et al., 1977]. Subsequently, developments in these technologies allow the routine sequencing of organismal genomes, metagenomes and marker genes from all domains of life. Genomic information can be seen as the 'blueprint' of life and being able to decode and to interpret it, grants insight into life's fundamental mechanisms [Moxon and Higgins, 1997,Henry et al., 2010]. However, microbes pose a challenge to genomic description as the vast majority of microbial life cannot readily be isolated in pure cultures [Amann et al., 1995,Curtis et al., 2002]. The rise of cultivation independent approaches like metagenomic and sequencing of marker genes addresses this limitation [Handelsman, 2004]. In these approaches, bulk DNA is extracted from an environmental sample and either specific genes are amplified and sequenced or random sequencing is performed. Thus, a fragmented, but cultivation-independent, overview of an environment's biological diversity and functional potential is provided [Pruesse et al.,

**2007, Ratnasingham and Hebert, 2007].**

**Early on, scientists recognized the necessity to share sequence data to facilitate reuse, reproducibility and comparisons. This has become an integral part of the research and publication process. In the 'Bermuda Principles', on the first international strategy meeting on human genome sequencing in 1996, it was agreed upon, that all human genomic sequence information, generated by centers funded for large-scale human sequencing, should be freely available in the public domain to encourage research and to maximize its benefits to society. In the Fort Lauderdale meeting in 2003 organized by the Wellcome Trust, it was finally agreed to deposit all kinds of sequencing data that are analyzed in scientific publications in public databases. Over the past two decades, the amount of sequence data submitted to the world's largest public nucleotide sequence data repository INSDC (International Nucleotide Sequence Database Collaboration, comprising of DDBJ (DNA Data Bank of Japan), ENA (European Nucleotide Archive), and GenBank) has grown exponentially [Stratton et al., 2009]. Recently, Next Generation Sequencing (NGS) technologies [Mardis, 2008] allow even faster and more economical sequence generation, resulting in an unprecedented sequence accumulation.**

**Despite the impressive magnitude of sequence data generation, numerous life science studies have shown that contextual (meta)data (CD) are crucial for their interpretation [DeLong et al., 2006, Fuhrman et al., 2006, Schriml et al., 2010]. CD are metadata about features such as the environmental origin and the processing steps that were applied to obtain the sequences. These ranges from data about the geographic location (latitude, longitude), sampling time, habitat, to experimental procedures used to obtain the sequences up to video data recorded during sampling. The fact however that e.g. latitude, longitude (INSDC: lat_lon), and time (INSDC: collection_date), which can be submitted to the public repositories since years, have so far only been reported in 7.3% and 7.2% of all submissions [Hankeln et al., 2010], strongly implies that the procedure to deposit these data is hampered. Common reasons are: 1) no clear descriptors exist to guide the submitters which metadata should be deposited and 2) no appropriate tools exist that support the combined submission of sequence data and CD.**

**These concerns have recently prompted the Genomic Standards Con-**

**sortium (GSC), an international consortium, which promotes mechanisms to standardize the description of genomes and the exchange of genomic data, to create a series of checklists defining the minimal set of CD that should accompany a sequence submission. The Minimum Information About a (Meta)Genome Sequence (MIGS/MIMS) checklist [Field et al., 2008] outlines a conceptual structure for extending the core information that has been traditionally captured by the INSDC (DDBJ/EMBL/GenBank) to describe genomic and metagenomic sequences. The Minimum Information about a MARKer gene Sequence (MIMARKS) standard complements the MIGS/MIMS specification by adding two new "report types", a "MIMARKS-survey" and a "MIMARKS-specimen", the former being the checklist for uncultured diversity marker gene surveys, the latter is designed for marker gene sequences obtained from any material identifiable via specimens. The standards also cover sets of measurements and observations describing particular habitats, termed "environmental packages". Collectively the MIGS/MIMS/MIMARKS standards are now called MIxS (Minimum Information about any (x) Sequence) (Yilmaz et al., The Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications, accepted).**

**Through collaboration with the GSC, the INSDC now offers the structures to store the data items specified in the GSC checklists. This facilitates an early integration of sequence data and CD. However, specialized tools to allow this integration for different user scenarios are needed.**

**The European Nucleotide Archive (ENA) provides an on-line submission system called Webin which contains prepared web forms for the submission of GSC compliant data. It shows all fields with descriptions, explanations and examples and does data validation in the forms (**`https://www.ebi.ac.uk/embl/genomes/submission/login.jsf`**, accessed: 16.03.2011). The Investigation Study Assay (ISA) Infrastructure offers a software suite that produces documents that can be submitted to the Sequence Read Archive (SRA) repository [Rocca-Serra et al., 2010]. With the Quantitative Insights Into Microbial Ecology (QIIME) web application [Caporaso et al., 2010] users can generate and validate MIMARKS-compliant templates. Finally, MetaBar is a**

spreadsheet and web-based software tool which assists users in the consistent acquisition, electronic storage and submission of CD associated to their samples [Hankeln et al., 2010]. However, a tool that integrates CD and sequence data by directly enriching FASTA files for submission does not exist yet.

Here we present CDinFusion (Contextual Data and FASTA in fusion). CDinFusion has been designed to submit sequence data together with CD to INSDC. CDinFusion intends to facilitate the integration of CD and sequence data prior to submission by directly enriching sequence data using the FASTA format. It generates submission-ready outputs for INSDC by implementing the MIxS standard defined by the GSC. CDinFusion processes single as well as MultiFASTA files, containing up to millions of sequences. It was successfully applied to several use cases. Example submissions to the INSDC can be accessed with the following accession numbers: JF681370, JF268327-JF268425 and Genome Project ID 63253. A public installation is hosted and maintained at the Max Planck Institute for Marine Microbiology, Bremen, Germany: `http://www.megx.net/cdinfusion`. The tool is easy to install and released under the LGPL 3 open source license to promote distribution in aid of increasing the quantity and quality of CD in the public repositories.

## 3.3 Results

CDinFusion has been designed as a web-based tool, which enables users to upload single or MultiFASTA files from single sequence to high-throughput analysis and enrich them with CD. After uploading the sequences, the user is requested to select the appropriate GSC checklist and environmental package. CD can be entered in the web forms or CSV templates can be downloaded, filled with CD off-line and uploaded. The CSV files help to store and share the data. The merged sequence and CD can be downloaded for subsequent submission to INSDC.

The implemented workflow covers the three typical scenarios of sequence submission to an INSDC database namely: 1) Enriching a single sequence with one CD set, 2) Enriching many sequences in a Mul-

**tiFASTA file with one CD set, and 3) Enriching subsets of sequences in a MultiFASTA file with several CD sets (Figure 3.1). The functionality of each of these different scenarios has been tested in dedicated use cases. The first use case was conducted with a single 16S rRNA sequence obtained from a bacterium isolated from a coastal water sample taken off the coast of the Wadden Sea island Sylt. After uploading the FASTA file the tool directly proceeded to the CD package selection for one CD set, as the file contained only a single sequence. The MIMARKS survey (mimarks_s) package and the water package were selected to provide suitable CD fields for this environmental survey**

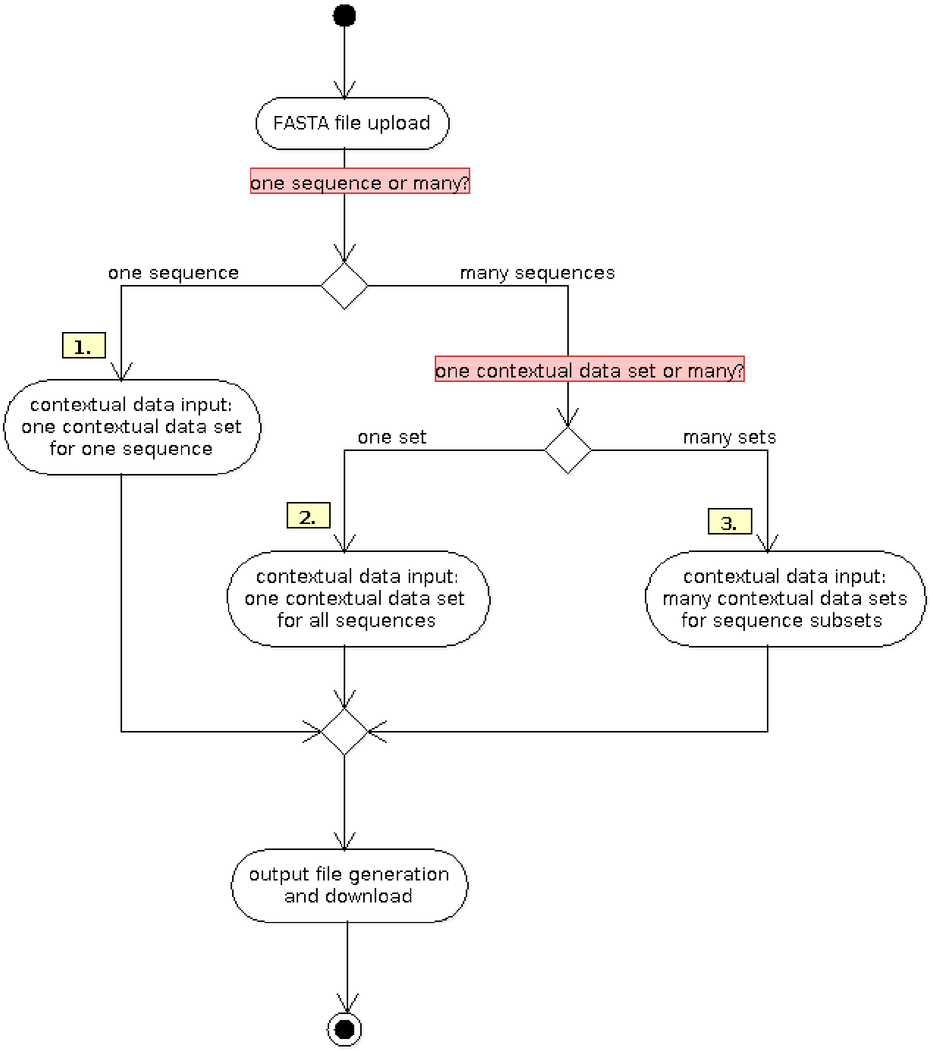
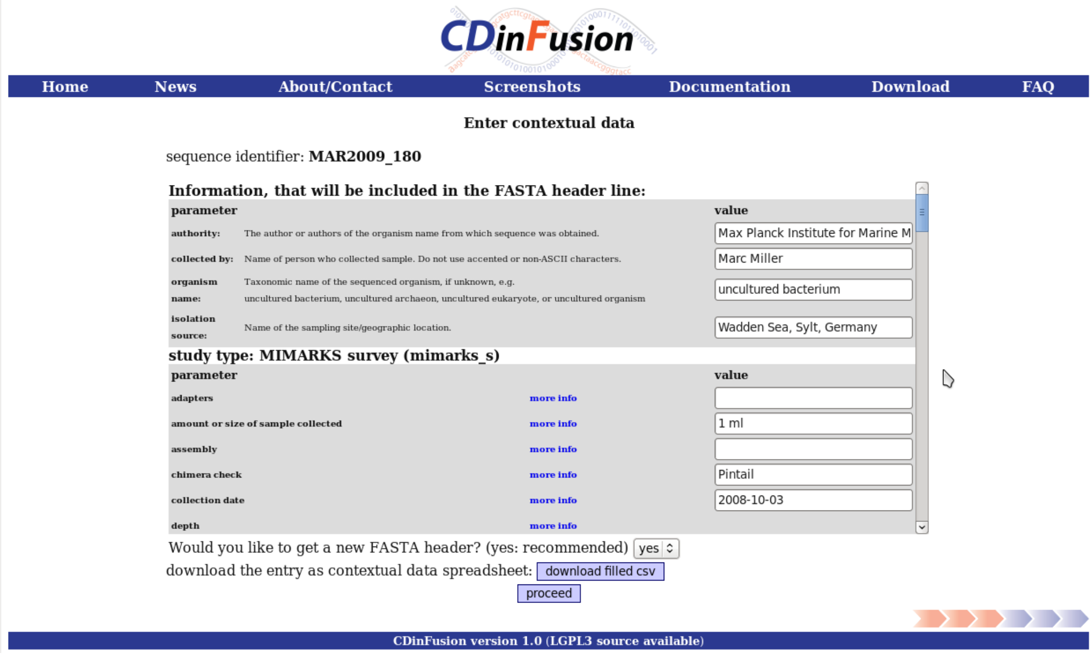

FASTA file upload
one sequence or many?
one sequence
many sequences
1.
contextual data input: one contextual data set for one sequence
one contextual data set or many?
one set
many sets
2.
3.
contextual data input: one contextual data set for all sequences
contextual data input: many contextual data sets for sequence subsets
output file generation and download

Figure 3.1: Overview of submission scenarios. Three primary scenarios of sequence data submission to INSDC can be distinguished and are all covered by the CDinFusion workflow: 1) The submission of a single FASTA sequence file along with one CD set, 2) The submission of a MultiFASTA file along with one CD set for all sequences in the file and 3) The submission of a MultiFASTA file annotated with several CD sets.

sequence obtained from seawater. Subsequently the web forms were filled with all the CD available for this particular sequence (example Figure 3.2). After generating and downloading the output file, the CD enriched FASTA was imported into Sequin version 11.00. CDinFusion inserted qualifiers specified by GenBank into the header line of the FASTA file. The tool placed the rest of the CD into a tab delimited structured comment file. This file was loaded into Sequin with the "Advanced Table Readers" option in the "Annotate" menu. The CD appeared in the metadata section between the header and the feature table section. By selecting "Done", the Sequin file was saved and the complete submission was prepared. The INSDC database entry for this submission can be accessed at [Accession number: JF681370].



Enter contextual data

sequence identifier: **MAR2009_180**

**Information, that will be included in the FASTA header line:**

| parameter | | value |
|---|---|---|
| authority: | The author or authors of the organism name from which sequence was obtained. | Max Planck Institute for Marine M |
| collected by: | Name of person who collected sample. Do not use accented or non-ASCII characters. | Marc Miller |
| organism name: | Taxonomic name of the sequenced organism, if unknown, e.g. uncultured bacterium, uncultured archaeon, uncultured eukaryote, or uncultured organism | uncultured bacterium |
| isolation source: | Name of the sampling site/geographic location. | Wadden Sea, Sylt, Germany |

**study type: MIMARKS survey (mimarks_s)**

| parameter | | value |
|---|---|---|
| adapters | more info | |
| amount or size of sample collected | more info | 1 ml |
| assembly | more info | |
| chimera check | more info | Pintail |
| collection date | more info | 2008-10-03 |
| depth | more info | |

Would you like to get a new FASTA header? (yes: recommended) yes

download the entry as contextual data spreadsheet: download filled csv

Figure 3.2: CDinFusion web user interface. The CD are entered into the auto-generated web forms. Details about each parameter are accessible with the "more info" link. These details are retrieved using a web service accessing the GSC database and are therefore always up to date.

This use case exemplifies submission scenarios, where a single sequence and its CD are to be submitted to the INSDC databases. Single sequences can, for example, be marker genes or genomes that consist of a single sequence or contig.

In the second use case, a permanent draft genome from a ***Rhodopirellula baltica*** strain along with its associated CD was prepared for submission. After the 6.9 Mb MultiFASTA file was uploaded, the user was offered the option to annotate all sequences in this file with one CD set or to enter many CD sets for sequence subsets. As all sequence fragments were parts of the same bacterial genome, isolated from a sediment sample, one CD set for all sequences was selected using the MIGS bacterial genome (ba) checklist and the sediment package. The user filled in all CD fields available and the CD enriched files were generated, downloaded and imported into Sequin. The data of this genome project can be accessed by ID 63253 and with the accession number: AFAR00000000. The genome will be analyzed in a separate study in preparation (Richter et al., Permanent draft genome sequence of *Rhodopirellula baltica* WH47).

This use case describes a procedure that may also be applied to metagenomic MultiFASTA files originating from one sampling site, which should be annotated with the same CD.

In the third use case a MultiFASTA file containing 99 16S rRNA sequences, obtained from a clone library, was enriched with CD. This file comprised four sequence subgroups, each with distinct CD. After the MultiFASTA file was uploaded, the CD for each of the groups was entered sequentially until all sequence subgroups were annotated. After the user selected the MIMARKS (mimarks_s) and the "environmental package" sediment the CD were entered in the web forms.

The output files created were a CD enriched MultiFASTA file and a compressed ZIP archive containing four structured comment files, one for each of the subgroups. After the FASTA file had been imported to Sequin, the structured comment files were loaded one by one with the "Advanced Table Readers" function. The file was then saved and submitted. This clone library and its CD [Bialas et al., 2007] will be analyzed in a separate study in preparation (Ruff et al., Microbial Communities of Submarine Methane Seeps at Hikurangi Margin, New Zealand). The INSDC database entries for this submission will be available under the accession numbers: JF268327-JF268425. The same procedure has been applied to ten 16S rRNA sequences of an environmental culturability study conducted by the M.Sc. Marine Microbiology (MarMic) class of 2014 at the island of Sylt. The sequences of that study will be analyzed in a separate study in preparation (Hahnke et al., *Flavobacteria* of the North Sea: Diversity of Culturability) available under the accession numbers: JF710778-JF710788.

This use case applies, whenever batches of sequences have to be submitted and subgroups of these sequences have to be annotated with individual CD sets. These MultiFASTA files can for example contain batches of marker genes or a pooled metagenome.

To test if high-throughput data can be processed with CDinFusion, metagenomic FASTA files from the Global Ocean Survey (GOS, http://jcvi.org/cms/research/projects/gos/overview/), and metagenome data from the Microbial Interactions in Marine Systems project (MIMAS, http://www.mimas-projekt.de/mimas/, accessed: 16.03.2011) were loaded into CDinFusion. FASTA files containing several million sequences with file sizes of several GigaBytes (GB) could be processed in less than three minutes in an AMD$^{TM}$ 64Bit, 2 GHz and 4 GB RAM environment.

All described test cases were recorded with the Selenium IDE (http://seleniumhq.org/) test case recorder. The test cases along with the test data, except for the metagenomic datasets, are deposited at http://code.google.com/p/cdinfusion. Descriptions how to run the tests, can be found in the documentation section of the public CDinFusion installation at http://www.megx.net/cdinfusion.

## 3.4 Design and Implementation

### Languages, Tools and detailed Workflow

CDinFusion has been designed to allow users to add CD to single and MultiFASTA files that may comprise one to several million sequences. The CD enriched output can readily be submitted to the INSDC archives. The tool is programmed in the object-oriented, platform-independent programming language Java SE 5.0 (http://www.oracle.com/technetwork/java/index.html) using the Eclipse IDE (http://www.eclipse.org/). The open source Spring framework (http://www.spring-source.org/about/) was used, which supports the Model-View-Controller (MVC) design pattern. The functionality of the tool was continuously tested using the Selenium IDE (http://seleniumhq.org/). It runs on an Apache Tomcat 5.5.25 web server (http://tomcat.apache.org/). The project has been built using Apache Ant 1.7.1 (http://ant.apache.org/) and has been deployed on a web server with 2 GHz

**AMD Opteron$^{TM}$ processor 246, with 4 GB main memory and Debian GNU/Linux 5.0.3 (lenny).**

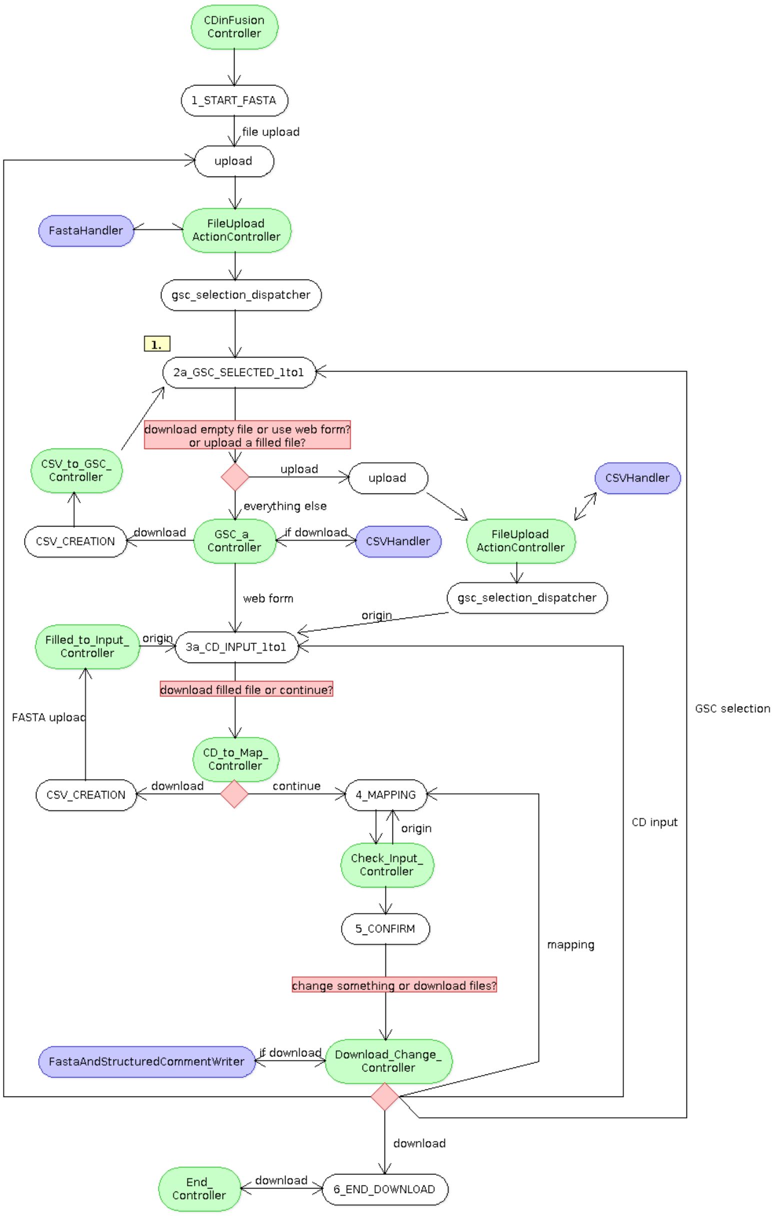


Figure 3.3: CDinFusion implementation details along the workflow 1, when a single sequence is submitted to the INSDC. CDinFusion implements the Model-View-Controller design pattern. Classes implementing the data model and its manipulation methods are shown in blue, components belonging to the web user interface (view) are shown in white and components directing the workflow (control) are shown in green.

**Figures 3.3 and 3.4 show the implementation details of the software's workflow. FASTA files are parsed and validated, when uploaded by the FastaReader class. It implements the FastaValidatorCallback interface**

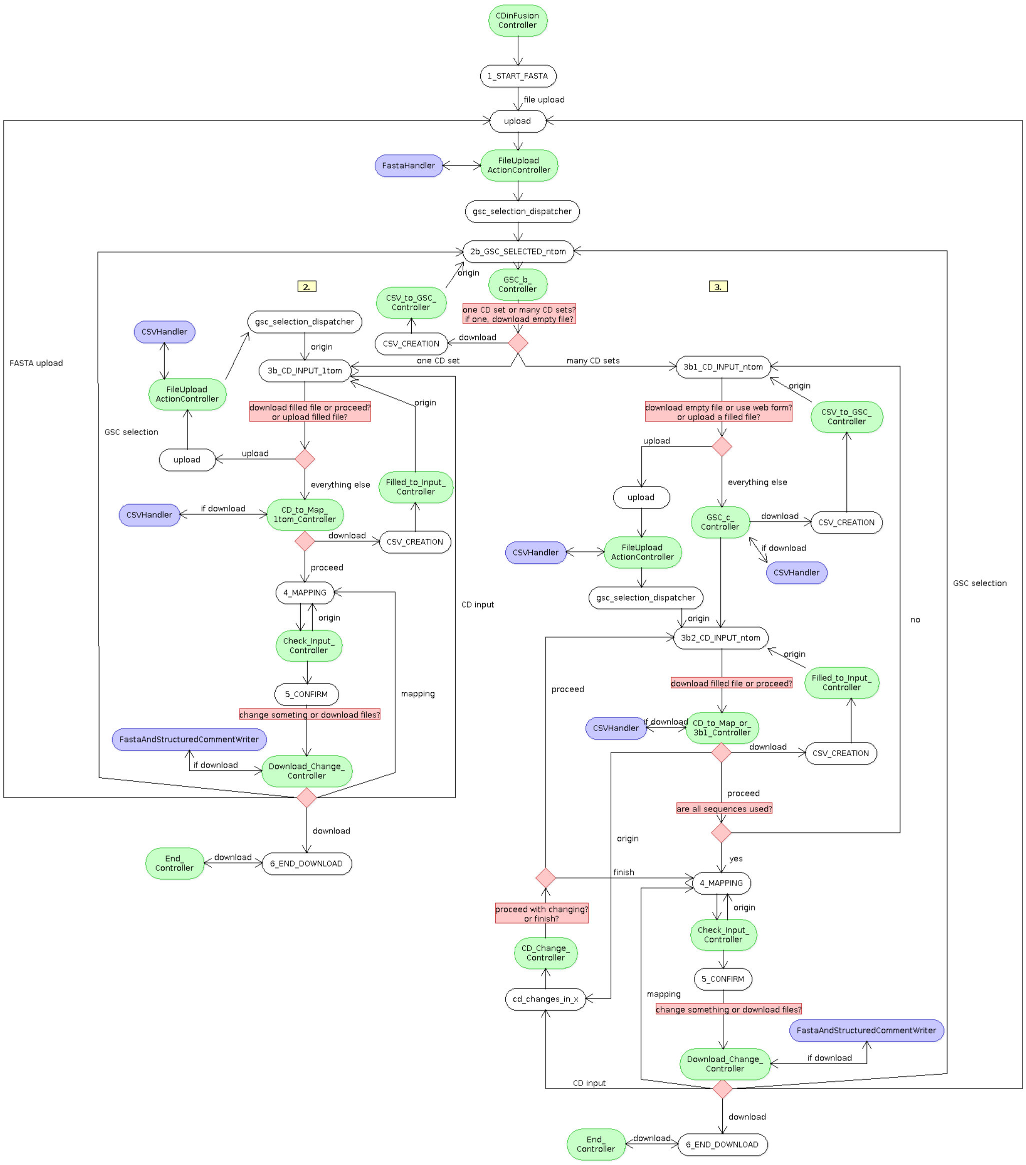


Figure 3.4: CDinFusion implementation details along the workflows 2 and 3 covering the primary scenarios of sequence data submission to the INSDC are shown. CDinFusion implements the Model-View-Controller design pattern. Classes implementing the data model and its manipulation methods are shown in blue, components belonging to the web user interface (view) are shown in white and components directing the workflow (control) are shown in green.

**of the FastaValidator package (`http://www.megx.net/FastaValidator`), which has been developed within the frame of this project. This event-driven parser is designed to quickly parse and validate arbitrarily large FASTA files with minimal time and memory requirements. It facilitates the processing of gigabases of FASTA files containing millions of sequences on common desktop PC architectures. The parser is avail-**

able separately and is also released under the GNU LGPL 3 license. It may also be used for other projects.

If only one sequence is detected in the FASTA upload, the control flow will be directed towards the 2a_GSC_SELECTED_1to1 JSP (use case 1 in the Results section), shown in Figure 3. If the user opts to annotate all sequences of a MultiFASTA file (Figure 3.4) with either one CD set or many CD sets, the control flow will be directed either to the 3b_CD_INPUT_1tom JSP (use case 2 in the Results section) or to the 3b1_CD_INPUT_ntom JSP (use case 3 in the Results section), respectively.

After the CD have been entered into the web forms, these data may be downloaded as comma separated value (CSV) files. The CSV files may serve as local backups and can be edited off-line and uploaded to CDinFusion to re-populate the web forms. Each session concludes with a confirmation step, where users can revisit any previous step and correct CD input if necessary. This holds true for all three branches of the workflow (Figure 3.3 and 3.4). If the user chooses to proceed to the file download, a CD FASTA file and a structured comment file are generated and can, depending on their size, either be imported to Sequin or merged on the command line using tbl2asn (`http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html`, accessed: 30.03.2011) before submission.

### Implementation of the GSC checklists in CDinFusion

Once the user has uploaded a MultiFASTA file and its contents have been validated, the data is processed along the data model shown in Figure 3.5. For each CD set a CDElement is created that contains an object for a "type of report" and an object for an "environmental package". These Java classes were auto-generated from the relations in the PostgreSQL GSC database using the Ibator tool from the iBatis project (`http://ibatis.apache.org`). The GSC database is hosted and maintained at the Max Planck Institute for Marine Microbiology, Bremen, Germany. The Java classes cover the MIGS, MIMS and MIMARKS (MIxS) specifications. The GSC plans to refine these standards annually. With every new version of the standards the Java classes can easily be updated using the Ibator tool.

The short names of the parameters are resolved using a web service

that was developed within the frame of this project. The web service offers details about all GSC parameters stored in the GSC database. Web forms (see Figure 3.2) are dynamically rendered during runtime and therefore always contain the latest information including all definitions and descriptions of the GSC checklists parameters. If a user wants to know how a certain GSC parameter is specified, the "more info" link opens a window with information about the full name of the parameter, its definition, the expected value, the syntax and an example. This information is directly retrieved from the GSC database. For CDinFusion to be fully functional, there needs to be Internet access to the web service. If a certain type of report and environmental package has been selected, these parameters are cached. The next time these packages are selected the web forms are rebuild from cache without re-using the web service.
Two Strings "first SequenceID" and "lastSequenceID" in the CDElement object store the range of the associated sequence identifiers for each CD set. The CDFastaHeader object contains those parameters that are covered by the web forms in addition to the GSC parameters that are later used to extend the FASTA header lines.

## Installation details

There are two ways to install CDinFusion: 1) CDinFusion can be installed by downloading and deploying the pre-compiled web archive file (war) on an Apache Tomcat (version > 5.5.25). In this case the war file only has to be uploaded in the Tomcat manager. Afterwards the application can be accessed under `http://<local_tomcat_installation>/CDinFusion`. This method is preferable if users do not want to compile the program from its source code. 2) CDinFusion can also be installed by downloading and compiling the source code and subsequently deploying the software on an Apache Tomcat web server (version > 5.5.25). To compile the code, the generic build.xml and build.properties files can be adjusted to local settings. If the standard settings in these files are not changed, the war file will be compiled into the CDinFusion root folder. The project can be compiled by executing the Apache ant build tasks, "deploy" or "deploywar", respectively. The build.xml can additionally be configured to directly deploy the tool on an Apache Tomcat web server or to create the war file

**and upload it with the Tomcat manager. Further installation details can be found in the README.txt file that is included in the source bundle and that is also available in the documentation section of the CDinFusion web page. On some platforms the CATALINA_HOME environment variable needs to be set, in order for CDinFusion to write and read files. Relative to the path specified, CDinFusion will create a "data" folder, where temporary files will be saved. The application has been tested on Debian GNU Linux installations, but should be platform-independent and run on all platforms that support Java and Apache Tomcat installation such as Windows$^{\text{TM}}$ or MAC OS$^{\text{TM}}$.**

## Availability and Future Directions

**The public installation of CDinFusion is hosted and maintained at the Microbial Genomics and Bioinformatics Group (MGG) of the Max Planck Institute of Marine Microbiology Bremen and accessible under:** `http://www.megx.net/cdinfusion`**. The source code is available under GNU LGPL 3 and deposited in a public repository:** `http://code.google.com/p/cdinfusion`**.**

**As open source software it is the intention of the MGG to support this software well into the future. Currently CDinFusion supports submission of CD enriched sequence data to the INSDC using Sequin and tbl2asn for large data sets. Support for installations outside the MPI cannot be granted. The direct submission to EMBL/ENA and DDBJ is planned. Furthermore the integration of GCDML [Kottmann et al., 2008] as an exchange format would be advantageous. The GSC and life science community is encouraged to download the source code and to modify and extend the software to make it even more useful.**

**Acknowledgments** **Thanks to S. Emil Ruff and the Geotechnologien project COMET/MUMM II (03G0608A, BMBF) for providing unpublished data and for the beta-testing. Thanks to Michael Richter for genome data and beta-testing. Thanks to Rudolf Amann, Bernhard Fuchs and Hanno Teeling for the MIMAS data. Thanks to Jens Harder, Richard Hahnke and the M.Sc. Marine Microbiology (MarMic) class of 2014 for data and beta-testing.**

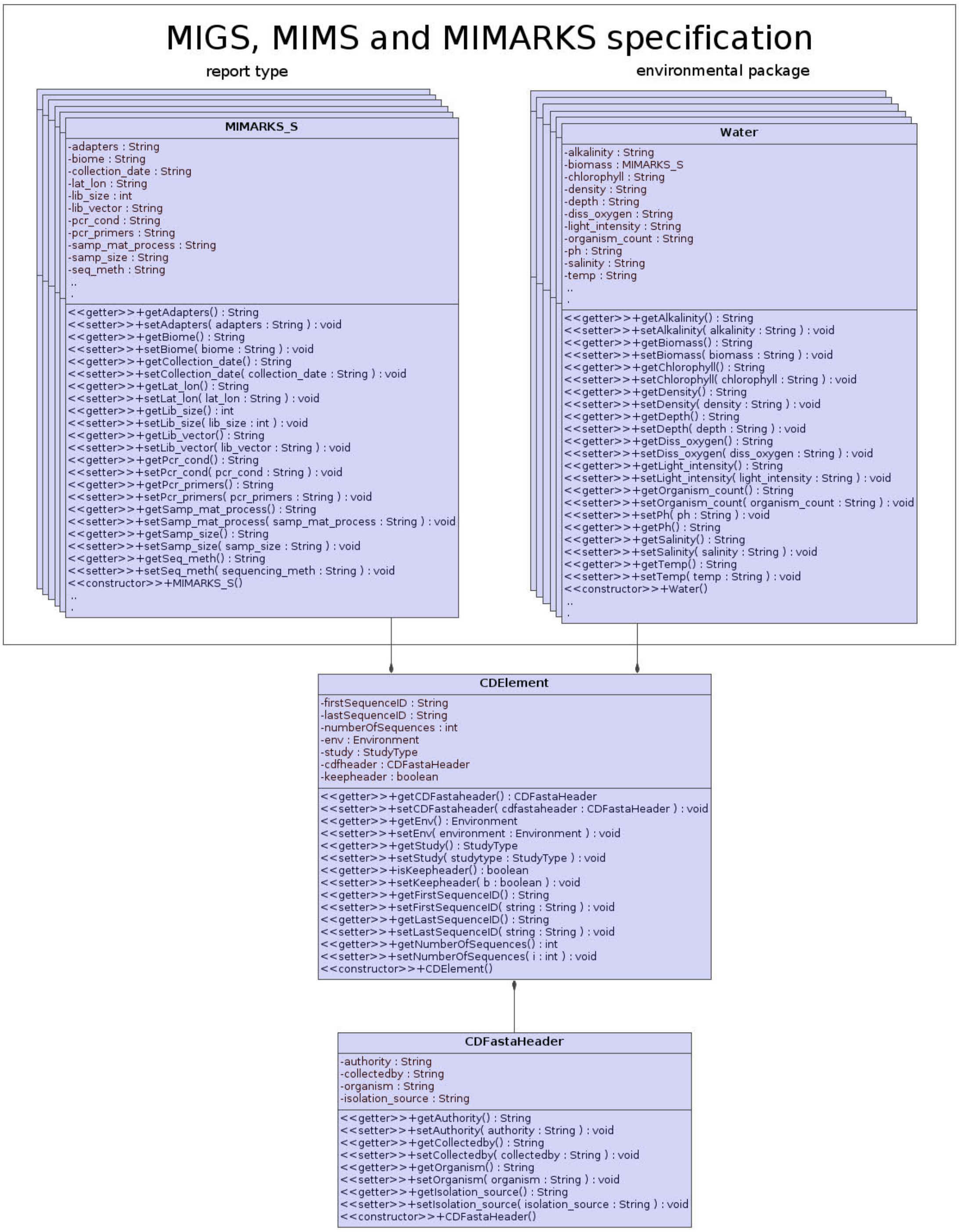

MIGS, MIMS and MIMARKS specification
report type
environmental package
MIMARKS_S
-adapters : String
-biome : String
-collection_date : String
-lat_lon : String
-lib_size : int
-lib_vector : String
-pcr_cond : String
-pcr_primers : String
-samp_mat_process : String
-samp_size : String
-seq_meth : String
<<getter>>+getAdapters() : String
<<setter>>+setAdapters( adapters : String ) : void
<<getter>>+getBiome() : String
<<setter>>+setBiome( biome : String ) : void
<<getter>>+getCollection_date() : String
<<setter>>+setCollection_date( collection_date : String ) : void
<<getter>>+getLat_lon() : String
<<setter>>+setLat_lon( lat_lon : String ) : void
<<getter>>+getLib_size() : int
<<setter>>+setLib_size( lib_size : int ) : void
<<getter>>+getLib_vector() : String
<<setter>>+setLib_vector( lib_vector : String ) : void
<<getter>>+getPcr_cond() : String
<<setter>>+setPcr_cond( pcr_cond : String ) : void
<<getter>>+getPcr_primers() : String
<<setter>>+setPcr_primers( pcr_primers : String ) : void
<<getter>>+getSamp_mat_process() : String
<<setter>>+setSamp_mat_process( samp_mat_process : String ) : void
<<getter>>+getSamp_size() : String
<<setter>>+setSamp_size( samp_size : String ) : void
<<getter>>+getSeq_meth() : String
<<setter>>+setSeq_meth( sequencing_meth : String ) : void
<<constructor>>+MIMARKS_S()
Water
-alkalinity : String
-biomass : MIMARKS_S
-chlorophyll : String
-density : String
-depth : String
-diss_oxygen : String
-light_intensity : String
-organism_count : String
-ph : String
-salinity : String
-temp : String
<<getter>>+getAlkalinity() : String
<<setter>>+setAlkalinity( alkalinity : String ) : void
<<getter>>+getBiomass() : String
<<setter>>+setBiomass( biomass : String ) : void
<<getter>>+getChlorophyll() : String
<<setter>>+setChlorophyll( chlorophyll : String ) : void
<<getter>>+getDensity() : String
<<setter>>+setDensity( density : String ) : void
<<getter>>+getDepth() : String
<<setter>>+setDepth( depth : String ) : void
<<getter>>+getDiss_oxygen() : String
<<setter>>+setDiss_oxygen( diss_oxygen : String ) : void
<<getter>>+getLight_intensity() : String
<<setter>>+setLight_intensity( light_intensity : String ) : void
<<getter>>+getOrganism_count() : String
<<setter>>+setOrganism_count( organism_count : String ) : void
<<setter>>+setPh( ph : String ) : void
<<getter>>+getPh() : String
<<getter>>+getSalinity() : String
<<setter>>+setSalinity( salinity : String ) : void
<<getter>>+getTemp() : String
<<setter>>+setTemp( temp : String ) : void
<<constructor>>+Water()
CDElement
-firstSequenceID : String
-lastSequenceID : String
-numberOfSequences : int
-env : Environment
-study : StudyType
-cdfheader : CDFastaHeader
-keepheader : boolean
<<getter>>+getCDFastaheader() : CDFastaHeader
<<setter>>+setCDFastaheader( cdfastaheader : CDFastaHeader ) : void
<<getter>>+getEnv() : Environment
<<setter>>+setEnv( environment : Environment ) : void
<<getter>>+getStudy() : StudyType
<<setter>>+setStudy( studytype : StudyType ) : void
<<getter>>+isKeepheader() : boolean
<<setter>>+setKeepheader( b : boolean ) : void
<<getter>>+getFirstSequenceID() : String
<<setter>>+setFirstSequenceID( string : String ) : void
<<getter>>+getLastSequenceID() : String
<<setter>>+setLastSequenceID( string : String ) : void
<<getter>>+getNumberOfSequences() : int
<<setter>>+setNumberOfSequences( i : int ) : void
<<constructor>>+CDElement()
CDFastaHeader
-authority : String
-collectedby : String
-organism : String
-isolation_source : String
<<getter>>+getAuthority() : String
<<setter>>+setAuthority( authority : String ) : void
<<getter>>+getCollectedby() : String
<<setter>>+setCollectedby( collectedby : String ) : void
<<getter>>+getOrganism() : String
<<setter>>+setOrganism( organism : String ) : void
<<getter>>+getIsolation_source() : String
<<setter>>+setIsolation_source( isolation_source : String ) : void
<<constructor>>+CDFastaHeader()

Figure 3.5: Data model. The central Java class is the CDElement class, which is a composition of of the classes "report type" and "environmental package". These classes implement the MIGS, MIMS and MIMARKS (MIxS) checklists specified by the GSC. The two strings "firstSequenceID" and "lastSequenceID" define if the CDElement contains CD for a single or a range of sequences. Instances of the CDFastaHeader class contain the data that is generated into the FASTA headers in the FASTA file.

CHAPTER 4

# MIMARKS

## The Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

**Authors:** Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison, Philippe Rocca-Serra, Peter Sterk, Manimozhiyan Arumugam, Mark Bailey, Laura Baumgartner, Bruce W Birren, Martin J Blaser, Vivien Bonazzi, Tim Booth, Peer Bork, Frederic D Bushman, Pier Luigi Buttigieg, Patrick S G Chain, Emily Charlson, Elizabeth K Costello, Heather Huot-Creasy, Peter Dawyndt, Todd DeSantis, Noah Fierer, Jed A Fuhrman, Rachel E Gallery, Dirk Gevers, Richard A Gibbs, Inigo San Gil, Antonio Gonzalez, Jeffrey I Gordon, Robert Guralnick, Wolfgang Hankeln, Sarah Highlander, Philip Hugenholtz, Janet Jansson, Andrew L Kau, Scott T Kelley, Jerry Kennedy, Dan Knights, Omry Koren, Justin Kuczynski, Nikos Kyrpides, Robert Larsen, Christian L Lauber, Teresa Legg, Ruth E Ley, Catherine A Lozupone, Wolfgang Ludwig, Donna Lyons, Eamonn Maguire, Barbara A Methé, Folker Meyer, Brian Muegge, Sara Nakielny, Karen E Nelson, Diana Nemergut, Josh D Neufeld, Lindsay K Newbold, Anna E Oliver, Norman R Pace, Giriprakash Palanisamy, Jörg Peplies, Joseph Petrosino, Lita Proctor, Elmar Pruesse, Christian Quast, Jeroen Raes, Sujeevan Ratnasingham, Jacques Ravel, David A Relman, Susanna Assunta-Sansone, Patrick D Schloss, Lynn Schriml, Rohini Sinha, Michelle I Smith, Erica Sodergren, Aymé Spor, Jesse Stombaugh, James M Tiedje, Doyle V Ward, George M Weinstock, Doug Wendel, Owen White, Andrew Whiteley, Andreas Wilke, Jennifer R Wortman, Tanya Yatsunenko and Frank Oliver Glöckner

**Submitted to:** nature biotechnology, accepted April 2011

**Personal Contribution:** Initial talk with the title: "Survey results: MInimal list of contextual data fields for ENvironmental Sequences (MIENS)" at the 6th meeting of the GSC at the EBI (Hinxton, UK) October 2008, which was the starting point for the development of this standard, that was later renamed to MIMARKS. Contributed suggestions for improvements of the data fields, during implementation work of this standard in the tools MetaBar and CDinFusion.

**Relevance:** Standards development for contextual data.

## 4.1 Abstract

Here we present a standard developed by the Genomic Standards Consortium (GSC) for reporting marker gene sequences—the minimum information about a marker gene sequence (MIMARKS). We also introduce a system for describing the environment from which a biological sample originates. The 'environmental packages' apply to any genome sequence of known origin and can be used in combination with MIMARKS and other GSC checklists. Finally, to establish a unified standard for describing sequence data and to provide a single point of entry for the scientific community to access and learn about GSC checklists, we present the minimum information about any (x) sequence (MIxS). Adoption of MIxS will enhance our ability to analyze natural genetic diversity documented by massive DNA sequencing efforts from myriad ecosystems in our ever-changing biosphere.

## 4.2 Introduction

Without specific guidelines, most genomic, metagenomic and marker gene sequences in databases are sparsely annotated with the information required to guide data integration, comparative studies and knowl-

**edge generation. Even with complex keyword searches, it is currently impossible to reliably retrieve sequences that have originated from certain environments or particular locations on Earth—for example, all sequences from 'soil' or 'freshwater lakes' in a certain region of the world. Because public databases of the International Nucleotide Sequence Database Collaboration (INSDC; comprising DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (EBI-ENA) and GenBank (`http://www.insdc.org/`)) depend on author-submitted information to enrich the value of sequence data sets, we argue that the only way to change the current practice is to establish a standard of reporting that requires contextual data to be deposited at the time of sequence submission. The adoption of such a standard would elevate the quality, accessibility and utility of information that can be collected from INSDC or any other data repository.**

**The GSC has previously proposed standards for describing genomic sequences— the 'minimum information about a genome sequence' (MIGS) — and metagenomic sequences — the 'minimum information about a metagenome sequence' (MIMS) [Field et al., 2008]. Here we introduce an extension of these standards for capturing information about marker genes. Additionally, we introduce 'environmental packages' that standardize sets of measurements and observations describing particular habitats that are applicable across all GSC checklists and beyond [Taylor et al., 2008]. We define 'environment' as any location in which a sample or organism is found, e.g., soil, air, water, human-associated, plant-associated or laboratory. The original MIGS/MIMS checklists included contextual data about the location from which a sample was isolated and how the sequence data were produced. However, standard descriptions for a more comprehensive range of environmental parameters, which would help to better contextualize a sample, were not included. The environmental packages presented here are relevant to any genome sequence of known origin and are designed to be used in combination with MIGS, MIMS and MIMARKS checklists.**

**To create a single entry point to all minimum information checklists from the GSC and to the environmental packages, we propose an overarching framework, the MIxS standard (`http://gensc.org/gc_wiki/index.php/MIxS`). MIxS includes the technology-specific checklists from the previous MIGS and MIMS standards, provides a way of in-**

troducing additional checklists such as MIMARKS, and also allows annotation of sample data using environmental packages. A schematic overview of MIxS along with the MIxS environmental packages is shown in Figure 4.1.

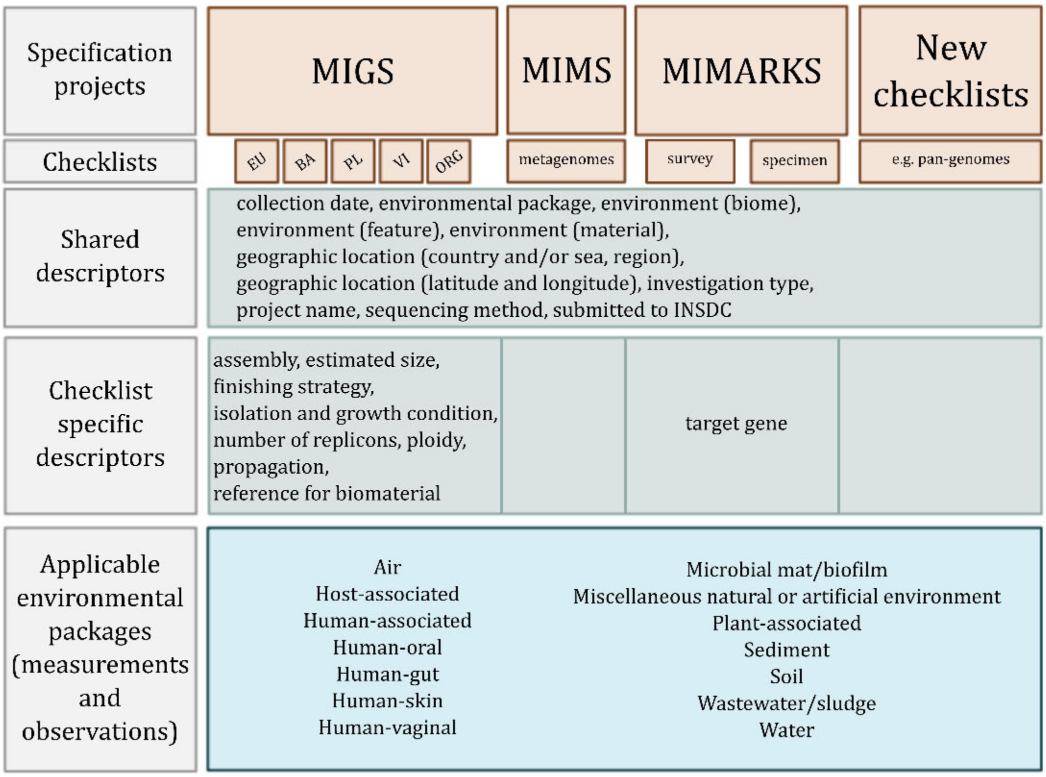

Specification projects
MIGS
MIMS
MIMARKS
New checklists
Checklists
EU
BA
PL
VI
ORG
metagenomes
survey
specimen
e.g. pan-genomes
Shared descriptors
collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC
Checklist specific descriptors
assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial
target gene
Applicable environmental packages (measurements and observations)
Air
Host-associated
Human-associated
Human-oral
Human-gut
Human-skin
Human-vaginal
Microbial mat/biofilm
Miscellaneous natural or artificial environment
Plant-associated
Sediment
Soil
Wastewater/sludge
Water

Figure 4.1: Schematic overview about the GSC MIxS standard (brown), including combination with specific environmental packages (blue). Shared descriptors apply to all MIxS checklists, however, each checklist has its own specific descriptors as well. Environmental packages can be applied to any of the checklists. EU, eukarya; BA, bacteria/archaea; PL, plasmid; VI, virus; ORG, organelle.

## 4.3 Development of MIMARKS and the environmental packages

Over the past three decades, the 16S rRNA, 18S rRNA and internal transcribed spacer gene sequences (ITS) from Bacteria, Archaea and microbial Eukaryotes have provided deep insights into the topology of the tree of life [Ludwig and Schleifer, 2005, Ludwig et al., 1998] and the composition of communities of organisms that live in diverse environments, ranging from deep sea hydrothermal vents to ice sheets in the Arctic [Giovannoni et al., 1990, Stahl et al., 1984, Ward et al., 1990, DeLong, 1992, Díez et al., 2001, Fuhrman et al., 1992, Hewson and Fuhrman, 2004, Huber et al., 2002, López-García et al., 2001, van der Staay et al., 2001, Pace, 1997, Rappé and Giovannoni, 2003]. Numerous other phylogenetic marker genes have proven useful, including RNA polymerase subunits (rpoB), DNA gyrases (gyrB), DNA recombination and repair proteins (recA) and heat shock proteins (HSP70) [Ludwig and Schleifer, 2005]. Marker genes can also reveal key metabolic functions rather than phylogeny; examples include nitrogen cycling (amoA, nifH, ntcA) [Francis et al., 2007, Zehr et al., 1998] sulfate reduction (dsrAB) [Minz et al., 1999] or phosphorus metabolism (phnA, phnI, phnJ) [Gilbert et al., 2009, Martinez et al., 2010]. In this paper we define all phylogenetic and functional genes (or gene fragments) used to profile natural genetic diversity as ‘marker genes’.

MIMARKS (Table 4.1) complements the MIGS/MIMS checklists for genomes and metagenomes by adding two new checklists, a MIMARKS survey, for uncultured diversity marker gene surveys, and a MIMARKS specimen, for marker gene sequences obtained from any material identifiable by means of specimens. The MIMARKS extension adopts and incorporates the standards being developed by the Consortium for the Barcode of Life (CBOL) [Hanner, 2009]. Therefore, the checklist can be universally applied to any marker gene, from small subunit rRNA to cytochrome oxidase I (COI), to all taxa, and to studies ranging from single individuals to complex communities.

Both MIMARKS and the environmental packages were developed by collating information from several sources and evaluating it in the framework of the existing MIGS/MIMS checklists. These include four independent community-led surveys, examination of the parameters reported in published studies, and examination of compliance with optional features in INSDC documents. The overall goal of these activities was to design the backbone of the MIMARKS checklist, which describes the most important aspects of marker gene contextual data.

## Results of community-led surveys

Four online surveys about descriptors for marker genes have been conducted to determine researcher preferences for core descriptors.

The Department of Energy Joint Genome Institute and SILVA [Pruesse

| Item | Description | Report type | |
|---|---|---|---|
| | | MIMARKS survey | MIMARKS specimen |
| **Investigation**<br>Submitted to INSDC$^{[boolean]}$ | Depending on the study (large-scale e.g., done with next-generation sequencing technology, or small-scale) sequences have to be submitted to SRA (Sequence Read Archives), DRA (DDBJ Sequence Read Archive) or through the classical Webin/Sequin systems to GenBank, ENA and DDBJ | M | M |
| Investigation type$^{[mimarks-survey]}$ or $[mimarks-specimen]$ | Nucleic Acid Sequence Report is the root element of all MIMARKS compliant reports as standardized by Genomic Standards Consortium (GSC). This field is either MIMARKS survey or MIMARKS specimen | M | M |
| Project name | Name of the project within which the sequencing was organized | M | M |
| **Environment**<br>Geographic location (latitude and longitude) $[float, point, transect, region]$ | The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system | M | M |
| Geographic location (depth $[integer, point, interval, unit]$) | Please refer to the definitions of depth in the environmental packages | E | E |
| Geographic location (elevation of site $^{[integer, unit]}$; altitude of sample $^{[integer, unit]}$) | Please refer to the definitions of either altitude or elevation in the environmental packages | E | E |
| Geographic location (country and/or sea $^{[INSDC]}$ or $^{[GAZ]}$; region $^{[GAZ]}$) | The geographical origin of the sample as defined by the country or sea name. Country, sea or region names should be chosen from the INSDC list (http://insdc.org/country.html), or the GAZ (Gazetteer, v1.446) ontology (`http://bioportal.bioontology.org/visualize/40651`) | M | M |
| Collection date$^{[ISO8601]}$ | The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated, that is, all of these are valid times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008; Except for 2008-01 and 2008, all are ISO6801 compliant | M | M |
| Environment (biome $^{[EnvO]}$) | In environmental biome level are the major classes of ecologically similar communities of plants, animals and other organisms. Biomes are defined based on factors such as plant structures, leaf types, plant spacing and other factors like climate. Examples include desert, taiga, deciduous woodland or coral reef. Environment Ontology (EnvO) (v1.53) terms listed under environmental biome can be found at `<http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00000428>` | M | M |
| Environment (feature $^{[EnvO]}$) | Environmental feature level includes geographic environmental features. Examples include harbor, cliff or lake. EnvO (v1.53) terms listed under environmental feature can be found at `<http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00002297>` | M | M |
| Environment (material $^{[EnvO]}$) | The environmental material level refers to the matter that was displaced by the sample, before the sampling event. Environmental matter terms are generally mass nouns. Examples include: air, soil or water. EnvO (v1.53) terms listed under environmental matter can be found at `<http://bioportal.bioontology.org/visualize/44405/?conceptid=ENVO%3A00010483>` | M | M |
| **MIGS/MIMS/MIMARKS Extension**<br>Environmental package [air, host-associated, human-associated, human-skin, human-oral, human-gut, human-vaginal, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated, sediment, soil, wastewater/sludge, water] | MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained. All environmental packages listed here are further defined in separate subtables. By giving the name of the environmental package, a selection of fields can be made from the subtables and can be reported | M | M |
| **Nucleic acid sequence source**<br>Isolation and growth conditions$^{[PMID, DOIorURL]}$ | Publication reference in the form of PubMed ID (PMID), digital object identifier (DOI) or URL for isolation and growth condition specifications of the organism/material | - | M |
| **Sequencing**<br>T arget gene or locus (e.g., 16S rRNA, 18S rRNA, nif, amoA, rpo) | Targeted gene or locus name for marker gene study | M | M |
| Sequencing method (e.g., dideoxy-sequencing, pyrosequencing, polony) | Sequencing method used; e.g., Sanger, pyrosequencing, ABI-solid | M | M |

Table 4.1: Items for the MIMARKS specification and their mandatory (M), conditionally mandatory (C) (the item is mandatory only when applicable to the study) or recommended (X) status for both MIMARKS-survey and MIMARKS-specimen checklists. Furthermore, "–" denotes that an item is not applicable for a given checklist. E denotes that a field has environment-specific requirements. For example, whereas "depth" is mandatory for the environments water, sediment or soil, it is optional for human-associated environments. MIMARKS-survey is applicable to contextual data for marker gene sequences, obtained directly from the environment, without culturing or identification of the organisms. MIMARKS-specimen, on the other hand, applies to the contextual data for marker gene sequences from cultured or voucher-identifiable specimens. Both MIMARKS-survey and specimen checklists can be used for any type of marker gene sequence data, ranging from 16S, 18S, 23S, 28S rRNA to COI, hence the checklists are universal for all three domains of life. Item names are followed by a short description of the value of the item in parentheses and/or value type in brackets as a superscript. Whenever applicable, value types are chosen from a controlled vocabulary (CV), or an ontology from the Open Biological and Biomedical Ontologies (OBO) foundry (`http://www.obofoundry.org/`). This table only presents the very core of MIMARKS checklists, that is, only mandatory items for each checklist. Supplementary Results 2 contains all MIMARKS items, the tables for environmental packages in the MIGS/MIMS/MIMARKS extension and GenBank structured comment name that should be used for submitting MIMARKS data to GenBank. In case of submitting to EBI-ENA, the full names can be used.

et al., 2007] surveys focused on general descriptor contextual data for a marker gene, whereas the Ribosomal Database Project (RDP) [Cole et al., 2009] focused on prevalent habitats for rRNA gene surveys, and the Terragenome Consortium [Vogel et al., 2009] focused on soil metagenome project contextual data (Supplementary Results 1). The above recommendations were combined with an extensive set of contextual data items suggested by an International Census of Marine Microbes (ICoMM) working group that met in 2005. These collective resources provided valuable insights into community requests for contextual data items to be included in the MIMARKS checklist and the main habitats constituting the environmental packages.

## 4.4 Survey of published parameters

We reviewed published rRNA gene studies, retrieved from SILVA and the ICoMM database MICROBIS (The Microbial Oceanic Biogeographic Information System <http://icomm.mbl.edu/microbis>) to further supplement contextual data items that are included in the respective environmental packages. In total, 39 publications from SILVA and >40 ICoMM projects were scanned for contextual data items to constitute the core of the environmental package subtables (Supplementary Results 1). In a final analysis step, we surveyed usage statistics of INSDC source feature key qualifier values of rRNA gene sequences contained in SILVA (Supplementary Results 1). Notably, <10% of the 1.2 million 16S rRNA gene sequences (SILVA release 100) were associated with even basic information such as latitude and longitude, collection date or PCR primers.

## 4.5 The MIMARKS checklist

The MIMARKS checklist provides users with an ‘electronic laboratory notebook’ containing core contextual data items required for consistent reporting of marker gene investigations. MIMARKS uses the MIGS/MIMS checklists with respect to the nucleic acid sequence source and sequencing contextual data, but extends them with further experimental contextual data such as PCR primers and conditions, or target gene name. For clarity and ease of use, all items within the MIMARKS checklist are presented with a value syntax description, as well as a clear definition of the item. Whenever terms from a specific ontology are required as the value of an item, these terms can be readily found in the respective ontology browsers linked by URLs in the item definition. Although this version of the MIMARKS checklist does not contain unit specifications, we recommend all units to be chosen from and follow the International System of Units (SI) recommendations. In addition, we strongly urge the community to provide feedback regarding the best unit recommendations for given parameters. Unit standardization across data sets will be vital to facilitate comparative studies in future. An Excel version of the MIMARKS checklist is provided on the GSC web site (<http://gensc.org/gc_wiki/index.php/MIMARKS>).

### The MIxS environmental packages

Fourteen environmental packages provide a wealth of environmental and epidemiological contextual data fields for a complete description of sampling environments. The environmental packages can be combined with any of the GSC checklists (Table 4.1 and Supplementary Results 2). Researchers within The Human Microbiome Project [Turnbaugh et al., 2007] contributed the host-associated and all human packages. The Terragenome Consortium contributed sediment and soil packages. Finally, ICoMM, Microbial Inventory Research Across Diverse Aquatic Long Term Ecological Research Sites and the Max Planck Institute for Marine Microbiology contributed the water package. The MIMARKS working group developed the remaining packages (air, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated and wastewater/sludge). The package names describe high-level habitat terms in order to be exhaustive. The miscellaneous natural or artificial environment package contains a generic set of parameters, and is included for any other habitat that does not fall into the other thirteen categories. Whenever needed, multiple packages may be used for the description of the environment.

### Examples of MIMARKS-compliant data sets

Several MIMARKS-compliant reports are included in Supplementary

Results 3. These include a 16S rRNA gene survey from samples obtained in the North Atlantic, an 18S pyrosequencing tag study of anaerobic protists in a permanently anoxic basin of the North Sea, a pmoA survey from Negev Desert soils, a dsrAB survey of Gulf of Mexico sediments and a 16S pyrosequencing tag study of bacterial diversity in the western English Channel (SRA accession no. SRP001108).

## Adoption by major database and informatics resources

Support for adoption of MIMARKS and the MIxS standard has spread rapidly. Authors of this paper include representatives from genome sequencing centers, maintainers of major resources, principal investigators of large- and small-scale sequencing projects, and individual investigators who have provided compliant data sets, showing the breadth of support for the standard within the community.

In the past, the INSDC has issued a reserved 'barcode' keyword for the CBOL7. Following this model, the INSDC has recently recognized the GSC as an authority for the MIxS standard and issued the standard with official keywords within INSDC nucleotide sequence records [Benson et al., 2008]. This greatly facilitates automatic validation of the submitted contextual data and provides support for data sets compliant with previous versions by including the checklist version as a keyword.

GenBank accepts MIxS metadata in tabular format using the sequin and tbl2asn submission tools, validates MIxS compliance, and reports the fields in the structured comment block. The EBI-ENA Webin submission system provides prepared web forms for the submission of MIxS compliant data; it presents all of the appropriate fields with descriptions, explanations and examples, and validates the data entered. One tool that can aid submitting contextual data is MetaBar [Hankeln et al., 2010], a spreadsheet and web-based software, designed to assist users in the consistent acquisition, electronic storage and submission of contextual data associated with their samples in compliance with the MIxS standard. The online tool CDinFusion (http://www.megx.net/cdinfusion) was created to facilitate the combination of contextual data with sequence data, and generation of submission-ready files.

The next-generation Sequence Read Archive (SRA) collects and displays MIxS-compliant metadata in sample and experiment objects. There are several tools that are already available or under development to assist users in SRA submissions. The myRDP SRA PrepKit allows users to prepare and edit their submissions of reads generated from ultra-high-throughput sequencing technologies. A set of suggested attributes in the data forms assist researchers in providing metadata conforming to checklists such as MIMARKS. The Quantitative Insights Into Microbial Ecology (QIIME) web application (http://www.microbio.me/qiime) allows users to generate and validate MIMARKS-compliant templates. These templates can be viewed and completed in the users' spreadsheet editor of choice (e.g., Microsoft Excel). The QIIME web-platform also offers an ontology lookup and geo-referencing tool to aid users when completing the MIMARKS templates. The Investigation/Study/Assay (ISA) is a software suite that assists in the curation, reporting and local management of experimental metadata from studies using one or a combination of technologies, including high-throughput sequencing [Rocca-Serra et al., 2010]. Specific ISA configurations (<http://isa-tools.org/tools.html>) have been developed to ensure MIxS compliance by providing templates and validation capability. Another tool, ISAconverter, produces SRA.xml documents, facilitating submission to the SRA repository. MIxS checklists are also registered with the BioSharing catalog of standards (<http://biosharing/org/>), set to progressively link minimal information specifications to the respective exchange formats, ontologies and compliant tools. Further detailed guidance for submission processes can be found under the respective wiki pages (http://gensc.org/gc_wiki/index.php/MIxS) of the standard.

## Maintenance of the MIxS standard

To allow further developments, extensions and enhancements of MIxS, we set up a public issue tracking system to track changes and accomplish feature requests (http://mixs.gensc.org/). New versions will be released annually. Technically, the MIxS standard, including MIMARKS and the environmental packages, is maintained in a relational database system at the Max Planck Institute for Marine Microbiology Bremen on behalf of the GSC. This provides a secure and stable mechanism for updating the checklist suite and versioning. In future, we plan to develop programmatic access to this database to allow automatic retrieval of the latest version of each checklist for INSDC databases and for GSC community resources. Moreover, the Genomic Contextual Data Markup Language is a reference implementation of the GSC checklists by the GSC and now implements the full range of MIxS standards. It is based on XML Schema technology and thus serves as an interoperable data exchange format for infrastructures based on web services [Kottmann et al., 2008].

## 4.6 Conclusions and call for action

The GSC is an international body with a stated mission of working towards richer descriptions of the complete collection of genomes and metagenomes through the MIxS standard. The present report extends the scope of GSC guidelines to marker gene sequences and environmental packages and establishes a single portal where experimentalists can gain access to and learn how to use GSC guidelines. The GSC is an open initiative that welcomes the participation of the wider community. This includes an open call to contribute to refinements of the MIxS standards and their implementations. The adoption of the GSC standards by major data providers and organizations, as well as the INSDC, supports efforts to contextually enrich sequence data and complements recent efforts to enrich other (meta)omics data. The MIxS standard, including MIMARKS, has been developed to the point that it is ready for use in the publication of sequences. A defined procedure for requesting new features and stable release cycles will facilitate implementation of the standard across the community. Compliance among authors, adoption by journals and use by informatics resources will vastly improve our collective ability to mine and integrate invaluable sequence data collections for knowledge- and application-driven research. In particular, the ability to combine microbial community samples collected from any source, using the universal tree of life as a measure to compare even the most diverse communities, should provide new insights into the dynamic spatiotemporal distribution of microbial life on our planet and on the human body.

***Note: Supplementary information is available on the Nature Biotechnology website.***

**Funding sources are listed in the Supplementary Note.**

## Competing financial interests

**The authors declare no competing financial interests.**

**Published online at** `http://www.nature.com/nbt/index.html`**.**

**Reprints and permissions information is available online at** `http://www.nature.com/reprints/index.html`**.**

CHAPTER 5

# MEGX.NET

## Integrated database resource for marine ecological genomics

**Authors:** Renzo Kottmann, Ivaylo Kostadinov, Melissa Beth Duhaime, Pier Luigi Buttigieg, Pelin Yilmaz, Wolfgang Hankeln, Frank Oliver Glöckner

**Published in:** Nucleic Acids Research, October 2010

**Personal Contribution:** Involvement in the web portal design and programming, bug fixing and improvement of the environmental data layers in the Genes Mapserver (with Renzo Kottmann, Ivaylo Kostadinov, Pier Luigi Buttigieg, Pelin Yilmaz).

**Relevance:** To enhance the core megx.net concept to the research platform of the Microbial Genomics Group at Max Planck Institute for Marine Microbiology.

## 5.1 Abstract

Megx.net is a database and portal that provides integrated access to georeferenced marker genes, environment data and marine genome and metagenome projects for microbial ecological genomics. All data are stored in the Microbial Ecological Genomics DataBase (MegDB), which is subdivided to hold both sequence and habitat data and global environmental data layers. The extended system provides access to several hundreds of genomes and metagenomes from prokaryotes and phages, as well as over a million small and large subunit ribosomal RNA sequences. With the refined Genes Mapserver, all data can be interactively visualized on a world map and statistics describing envi-

ronmental parameters can be calculated. Sequence entries have been curated to comply with the proposed minimal standards for genomes and metagenomes (MIGS/MIMS) of the Genomic Standards Consortium. Access to data is facilitated by Web Services. The updated megx.net portal offers microbial ecologists greatly enhanced database content, and new features and tools for data analysis, all of which are freely accessible from our webpage `http://www.megx.net`.

## 5.2 Introduction

Over the last years, molecular biology has undergone a paradigm shift, moving from a single experiment science to a high-throughput endeavour. Although the genomic revolution is rooted in medicine and biotechnology, it is currently the environmental sector, specifically the marine, which delivers the greatest quantity of data. Marine ecosystems, covering >70% of the Earth's surface, host the majority of biomass and significantly contribute to global organic matter and energy cycling. Micro-organisms are known to be the 'gatekeepers' of these processes and insights into their lifestyle and fitness will enhance our ability to monitor, model and predict future changes.

Recent developments in sequencing technology have made routine sequencing of whole microbial communities from natural environments possible. Prominent examples in the marine field are the ongoing Global Ocean Sampling (GOS) campaign [Venter et al., 2004, Rusch et al., 2007] and Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project (`http://www.moore.org/microgenome/`). Notably, the GOS resulted in a major input of new sequence data with unprecedented functional diversity [Yooseph et al., 2007]. The resulting flood of sequence data available in public databases is an extraordinary resource with which to explore microbial diversity and metabolic functions at the molecular level.

These large-scale sequencing projects bring new challenges to data management and software tools for assembly, gene prediction and annotation—fundamental steps in genomic analysis. Several new dedicated database resources have recently emerged to tackle the current need for large-scale metagenomic data management, namely CAM-

ERA [Seshadri et al., 2007], IMG/M [Markowitz et al., 2008] and MG-RAST [Meyer et al., 2008].

Nevertheless, it is increasingly apparent that the full potential of comparative genome and metagenome analysis can be achieved only if the geographic and environmental context of the sequence data is considered [Field et al., 2008, Field, 2008]. The metadata describing a sample's geographic location and habitat, the details of its processing, from the time of sampling to sequencing and subsequent analyses are important, e.g. modelling species' responses to environmental change or the spread and niche adaptation of bacteria and viruses. This suite of metadata is collectively referred as contextual data [Kottmann et al., 2008].

Megx.net is the first database to integrate curated contextual data with their respective genes, genomes and metagenomes in the marine environment [Lombardot et al., 2006]. Now, the extended megx.net database resource allows post factum retrieval of interpolated environmental parameters, such as temperature, nitrate, phosphate, etc. for any location in the ocean waters based on profile and remote sensing data. Furthermore, the content has been significantly updated to include prokaryote and marine phage genomes, metagenomes from the GOS project [Rusch et al., 2007] and all georeferenced small and large subunit ribosomal RNA (rRNA) sequences from the SILVA database project [Pruesse et al., 2007].

The extended megx.net portal is the first resource of its kind to offer access to this unique combination of data, including manually curated habitat descriptors for genomes, metagenomes and marker genes, their respective contextual data and additionally integrated environmental data. See the megx.net online video tutorial for a guided introduction and overview at `http://www.megx.net/portal/tutorial.html`.

## 5.3 New database structure and content

The Microbial Ecological Genomics DataBase (MegDB), the backbone of megx.net, is a centralized database based on the PostgreSQL database management system. The georeferenced data concerning geographic coordinates and time are managed with the PostGIS extension to PostgreSQL. PostGIS implements the 'Simple Features Specification for SQL' standard recommended by the Open Geospatial Consortium (OGC; http://www.opengeospatial.org/), and therefore offers hundreds of geospatial manipulation functions.

MegDB is comprised of (i) MetaStorage, which stores georeferenced DNA sequence data from a collection of genomes, metagenomes and genes of molecular environmental surveys, with their contextual data, and (ii) OceaniaDB, which stores georeferenced quantitative environmental data (Figure 5.1).

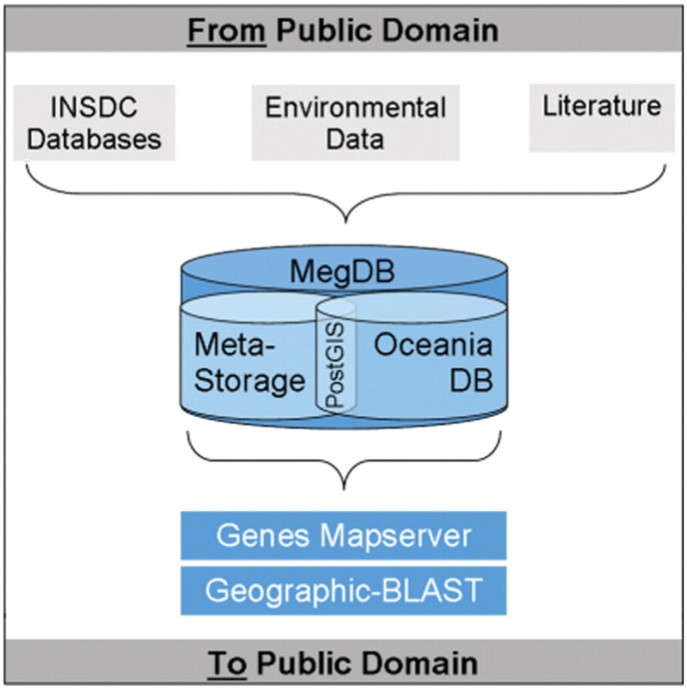


Figure 5.1: General architecture of megx.net: DNA sequence data (from INSDC) is integrated with contextual data from diverse resources (i.e. manual literature mining and the GOLD database) and interpolated environmental data. MegDB integrates the data conforming to OGC standards and MIGS/MIMS specification. The core megx.net tools, Genes Mapserver and Geographic-BLAST access the MegDB content.

## Contextual and sequence data content

Sequences in MetaStorage are retrieved from the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/). However, as of September 2009, GOLD reported 5776 genome projects, of which, only 1095 were finished and published (http://www.genomesonline.org/gold.cgi). As most of the sequenced functional diversity is contained in these draft and shotgun datasets, megx.net was extended to host draft genomes and whole genome shotgun data. Cur-

**rently, MegDB contains 1832 prokaryote genomes (940 incomplete or draft) and 80 marine shotgun metagenomes from the GOS microbial dataset. Marine viruses are a missing link in the correlation of microbial sequence data with contextual information to elucidate diversity and function. Consequently, megx.net now incorporates all sequenced marine phage genomes in MegDB, the first step towards a community call for integration of viral genomic and biogeochemical data [Brussaard et al., 2008].**

**In an effort towards integrating microbial diversity with specific sampling sites, megx.net has been extended to include georeferenced small and large subunit rRNA sequences from the SILVA rRNA databases project [Pruesse et al., 2007]. Currently, only 9% (16S/18S) and 2% (23S/28S) of over 1 million sequences in SILVA SSUParc (16S/18S) and LSUParc (23S/28S) databases are georeferenced. With the implementation of the Minimal Information about an Environmental Sequence (MIENS) standard for marker gene sequences (**`http://gensc.org/gc_wiki/index.php/MIENS`**), efforts are ongoing to significantly improve this situation.**

**All genomic sequences in megx.net are supplemented by contextual data from GOLD [Liolios et al., 2008] and NCBI Genome Projects (**`http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html`**). The database is designed to store all contextual data recommended by the Genomics Standards Consortium, and is thus compliant with the Minimum Information about a Genome Sequence (MIGS) standard and its extension, Minimum Information about a Metagenome Sequence (MIMS) [Field et al., 2008,Kottmann et al., 2008]. Furthermore, megx.net is the first resource to provide a manually annotated collection of genomes using terms from EnvO-Lite (Rev. 1.4), a subset of the Environment Ontology (EnvO) [Hirschman et al., 2008]. An EnvO-Lite term was assigned to each genome project, identifying the environment where its original sample material was obtained. The annotation can be browsed on the megx.net portal using, e.g. tag clouds, and may be used as a categorical variable in comparative analyses.**

### Environmental data content

**OceaniaDB was added to MegDB to supplement the georeferenced molecular data of MetaStorage with interpolated environmental pa-**

rameters. When sufficient date, depth and location measurements are provided, any 'on site' contextual data taken at a sampling site can be supplemented by environmental data describing physical, chemical, geological and biological parameters, such as ocean water temperature and salinity, nutrient concentrations, organic matter and chlorophyll.

The environmental data is retrieved from three sources:

1. World Ocean Atlas: a set of objectively analysed (one decimal degree spatial resolution) climatological fields of in situ measurements (`http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html`);

2. World Ocean Database: a collection of scientific, quality-controlled ocean profiles (`http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html`); and

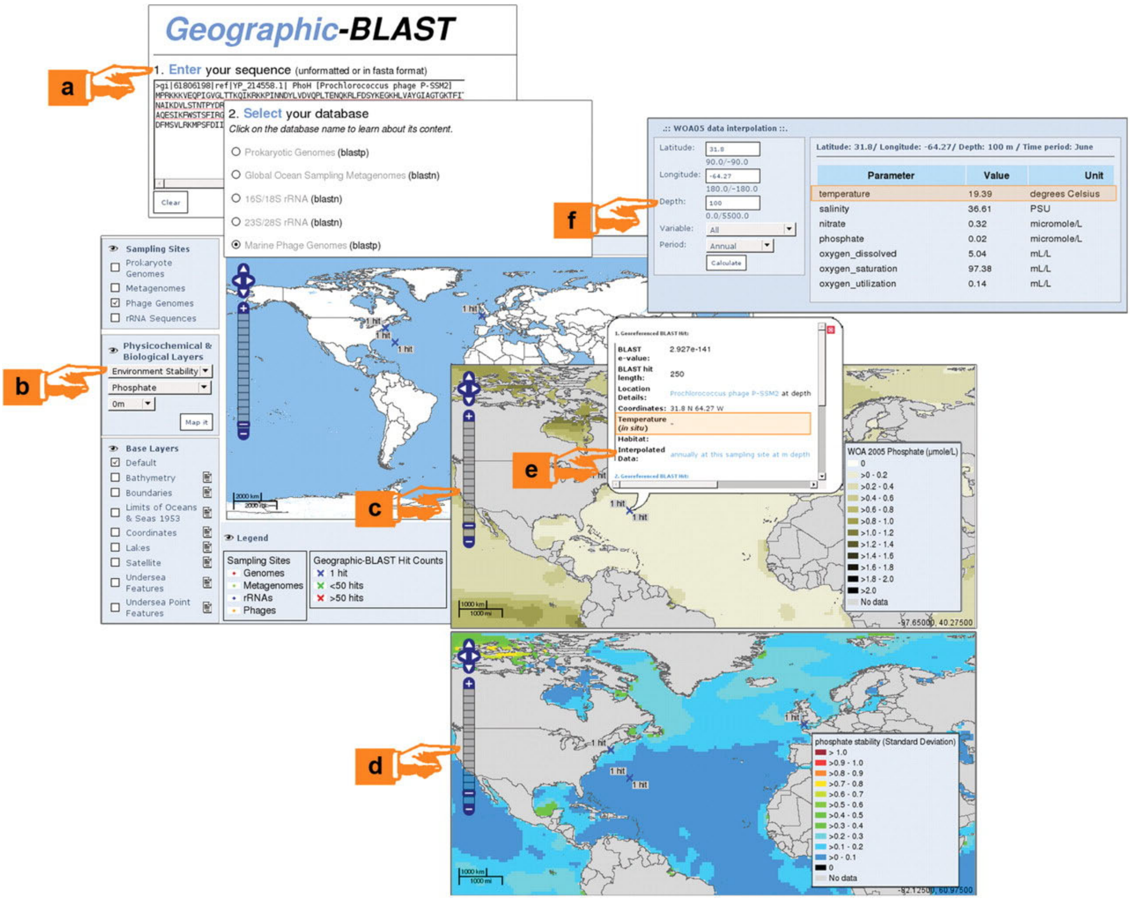
3. SeaWIFS chlorophyll a data (`http://seawifs.gsfc.nasa.gov`).

These data are described at 33 standard depths for annual, seasonal and monthly intervals. Together, the location and time data (x, y, z and t) serve as a universal anchor, and link environmental data to the sequence and contextual data in MetaStorage (Figure 5.1). As such, megx.net integrates biologist-supplied sequence and contextual data (measured at the time of sampling) with oceanographic data provided by third-party databases. All environmental data are compatible with OGC standards (`http://www.opengeospatial.org/standards`) and are described with exhaustive meta-information consistent with the ISO 19115 standard.
Moreover, based on the integrated environmental data, megx.net provides information to aid biologists in grasping the ocean stability, on both global and local scales. For all environmental parameters, the yearly standard deviations of the monthly values can be viewed on a world map, for easy visualization of high and low variation sample sites. Furthermore, for each sample site, users can view trends in numerous parameters.

## 5.4 User Access

### Genes Mapserver

**The Genes Mapserver (formerly Metagenomes Mapserver) offers a sample-centric view of the georeferenced MetaStorage content. Substantial improvements to the underlying Geographic Information System (GIS) and web view have been made. The website is now interactive, offering user-friendly navigation and an overlay of the OceaniaDB environmental data layers to display sampling sites on a world map in their environmental context. Sample site details and interpolated data can be retrieved by clicking the sampling points on the map (Figure 5.2). The GIS Tools of the Genes Mapserver allow extraction of inter-**



1. Enter your sequence (unformatted or in fasta format)

```
>gi|61806198|ref|YP_214558.1| PhoH [Prochlorococcus phage P-SSM2]
MPRKKKVEQPIGVGLTTKQIKRKKPINNDYLVDVQPLTENQKRLFDSYKEGKHLVAYGIAGTGKTFI
NAIKDVLSTNTPYDR
AQESIKFWSTSFIRG
DFMSVLRKMPSFDII
```

2. Select your database
*Click on the database name to learn about its content.*

○ Prokaryotic Genomes (blastp)
○ Global Ocean Sampling Metagenomes (blastn)
○ 16S/18S rRNA (blastn)
○ 23S/28S rRNA (blastn)
● Marine Phage Genomes (blastp)

.:: WOA05 data interpolation ::.

Latitude: 31.8 (90.0/-90.0)
Longitude: -64.27 (180.0/-180.0)
Depth: 100 (0.0/5500.0)
Variable: All
Period: Annual

Latitude: 31.8/ Longitude: -64.27/ Depth: 100 m / Time period: June

| Parameter | Value | Unit |
|---|---|---|
| temperature | 19.39 | degrees Celsius |
| salinity | 36.61 | PSU |
| nitrate | 0.32 | micromole/L |
| phosphate | 0.02 | micromole/L |
| oxygen_dissolved | 5.04 | mL/L |
| oxygen_saturation | 97.38 | mL/L |
| oxygen_utilization | 0.14 | mL/L |

1. Georeferenced BLAST Hit:

| | |
|---|---|
| BLAST e-value: | 2.927e-141 |
| BLAST hit length: | 250 |
| Location Details: | Prochlorococcus phage P-SSM2 at depth |
| Coordinates: | 31.8 N 64.27 W |
| Temperature (*in situ*) | - |
| Habitat: | |
| Interpolated Data: | annually at this sampling site at m depth |

2. Georeferenced BLAST Hit:

Figure 5.2: User test case: (a) BLAST sequence against the marine phage genomes to see the results on the Genes Mapserver. (b) View the BLAST hits with underlying environmental data, such as (c) average annual phosphate values, or (d) stability of phosphate concentrations in terms of monthly standard deviations. (e) BLAST result information can be displayed in a pop-up window, (f) where you can link out to megx.net's GIS data interpolator.

**polated values for several physicochemical and biological parameters,**

such as temperature, dissolved oxygen, nitrate and chlorophyll concentrations, over specified monthly, seasonally or annually intervals (Figure 5.2f).

### Geographic-BLAST

The Geographic-BLAST tool queries the MegDB genome, metagenome, marine phages and rRNA sequence data using the BLAST algorithm [Altschul et al., 1990]. The results are reported according to the sample locations (when provided) of the database hits. With the updated Geographic-BLAST, results are plotted on the Genes Mapserver world map, where they are labeled by number of hits per site (Figure 5.2). Standard BLAST results are shown in a table, which also provides direct access to the associated contextual data of the hits.

### Software extensions to the portal

In addition to the services directly provided by megx.net, the project serves as a portal to software for general data analysis in microbial genomics.

MetaBar (`http://www.megx.net/metabar`) is a tool developed with the aim to help investigators efficiently capture, store and submit contextual data gathered in the field. It is designed to support the complete workflow from the sampling event up to the metadata-enriched sequence submission to an INSDC database.

MicHanThi (`http://www.megx.net/michanthi`) is a software tool designed to facilitate the genome annotation process through rapid, high-quality prediction of gene functions. It clearly out-performs the human annotator in terms of accuracy and reproducibility.

JCoast (`http://www.megx.net/jcoast`; [Richter, 2003]) is a desktop application primarily designed to analyze and compare (meta)genome sequences of prokaryotes. JCoast offers a flexible graphical user interface, as well as an application programming interface that facilitates back-end data access to GenDB projects [Meyer et al., 2003]. JCoast offers individual, cross genome and metagenome analysis, including access to Geographic-BLAST.

### User test case

To demonstrate the interpretation of genomic content in environmental context, consider a test case with the marine phages. Marine phage genomes [Sullivan et al., 2005] and 'viral' classified GOS scaffolds [Williamson et al., 2008] have revealed host-related metabolic genes involved in, i.e. photosynthesis, phosphate stress, antibiotic resistance, nitrogen fixation and vitamin biosynthesis. Geographic-BLAST can be used to investigate the presence of PhoH (accession YP_214558), a phosphate stress response gene, among the sequenced marine phages. The search results can then be interpreted in their environmental context, either as (i) average annual phosphate measurements, or (ii) stability of phosphate concentrations in terms of monthly SD (Figure 2c and d). A closer look at a single genome sample site reveals that in situ temperature was not originally reported (Figure 2e), whereas the interpolated data supplements this parameter, among others (Figure 5.2f).

### Web Services

The newly extended version of megx.net offers programmatic access to MegDB content via Web Services, a powerful feature for experienced users and developers. All geographical maps can be retrieved via simple web requests, as specified by the Web Map Service (WMS) standard. The base URL for WMS requests is `http://www.megx.net/wms/gms`, where more detailed information on how to use this service can be found. Megx.net also provides access to MIGS/MIMS reports in Genomic Contextual Data Markup Language (GCDML) XML files for all marine phage genomes through similar HTTP queries, e.g. `http://www.megx.net/gcdml/Prochlorococcus_phage_P-SSP7.xml` [Field et al., 2008, Kottmann et al., 2008].

### Other changes

The massive influx of sequence data in the last years will out-compete the ability of scientists to analyze it [Anonymous, 2009]. This development already pushes megx.net's capability to provide comprehensive pre-computed data to the limit. To better focus on integration of molecular sequence, contextual and environmental data, megx.net no longer offers pre-computed analyses, especially considering that other facilities, such as MG-RAST and CAMERA have emerged. Furthermore, the 'EasyGenomes Browser' has been replaced with links to the

NCBI Genome

## 5.5 Summary

Since its first publication [Lombardot et al., 2006], megx.net has undergone extensive development. The web design has been revamped for better user experience, and the database content greatly enhanced, providing considerably more genomes and metagenomes, marine phages and rRNA sequence data.
Megx.net's unique integration of environmental and sequence data allows microbial ecologists and marine scientists to better contextualize and compare biological data, using, e.g. the Genes Mapserver and GIS Tools. The integrated datasets facilitate a holistic approach to understanding the complex interplay between organisms, genes and their environment. As such, megx.net serves as a fundamental resource in the emerging field of ecosystem biology, and paves the road to a better understanding of the complex responses and adaptations of organisms to environmental change.

### Database access

The database and all described resources are freely available at `http://www.megx.net/`. Continuously updated statistics of the content are available at `http://www.megx.net/content`. A web feed for news related to megx.net is available at `http://www.megx.net/portal/news/`. Feedback and comments, the most effective springboard for further improvements, are welcome at `http://www.megx.net/portal/contact.html` and via email to megx@mpi-bremen.de.

Overall, it is important to note that the megx.net website does not fully reflect the content and search functionalities of MegDB. For any specialized data request, contact the corresponding author.

### Funding

FP6 EU project MetaFunctions (CT 511784); Network of Excellence 'Marine Genomics Europe'; Max Planck Society. Funding for open

**access charge: Max Planck Society.**

## Conflict of interest statement.

**None declared.**

## Acknowledgements

**We would like to acknowledge Timmy Schweer, Thierry Lombardot, Magdalena Golden and Laura Sandrine for their valuable input to megx.net, as well as David E. Todd for redesigning the web page.**

CHAPTER 6

# DOMAINS OF UNKNOWN FUNCTION

## Ecological perspectives on domains of unknown function: a marine point of view

**Authors:** Pier Luigi Buttigieg, Wolfgang Hankeln, Melissa Beth Duhaime, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner

**Submitted to:** ISME Journal, January 2011

**Personal Contribution:** Helped to analyse the data after HMM calculations by generating network graphs together with Pier Luigi Buttigieg.

**Relevance:** Generating hypothesis about the possible functions of protein domains *in silico*, based on their co-occurrence patterns and environmental gradients.

## 6.1 Abstract

Metagenomic datasets from environmental samples offer attractive opportunities to characterize genomic elements of unknown function. We employed graph-theoretic approaches to visualize correlations between protein domains of unknown function detected in the Global Ocean Sampling metagenomes. Functional hypotheses for groups of these domains were generated based on network topology and existing putative functional assignments. Environmental contextualization of one such hypothesis was carried out using indirect gradient analysis.

## 6.2 Introduction

Genomic and metagenomic sequencing projects are revealing ever-increasing numbers of novel genes, many of unknown function. The Pfam 23 database [Finn et al., 2008], for example, stored some 10 340 protein domain families derived from conserved sequence data with 22% dubbed “domains of unknown function” (DUFs). This proportion is predicted to soon overtake that of functionally characterized domains [Bateman et al., 2010], prompting calls for community action [Roberts, 2004] and concerted, cross-disciplinary attention [Galperin and Koonin, 2010]. In their response, Jaroszewski et al. [Jaroszewski et al., 2009] and Goonesekere et al. [Goonesekere et al., 2010] noted several DUFs that appeared to be variations of functionally characterized protein folds, most likely maintained due to an extension of an organism’s ecological niche. It is reasonable to expect that conserved DUFs enhance ecological performance; however, characterizing DUFs from an ecological perspective has yet to be attempted. In this communication, we present a method of functional hypothesis generation based on DUF correlation across the Global Ocean Sampling (GOS) metagenome collection [Rusch et al., 2007]. Network visualizations were used in hypothesis generation followed by indirect gradient analysis to contextualize one well-defined hypothesis with environmental metadata. Together, these approaches aim to support efforts in DUF characterization using ecogenomic resources.

## 6.3 Material and Methods

Correlation analysis of microbial taxa and environmental parameters has previously been used to construct association networks [Fuhrman et al., 2008,Fuhrman, 2009]. Just as the correlation of taxa-abundance may elucidate a given taxon’s ecosystem-level interactions and function, correlation of protein domains across environments may grant insight into their potential associations and roles. This approach parallels the identification of unknown metabolic modules whereby genomic features found to co-vary in response to experimental perturbations are grouped in putative modules [Breitling et al., 2008]. To detect such associations in metagenomic datasets, we measured the Spearman rank correlation ($\rho$) between DUFs detected in the globally-distributed GOS metagenomes (473 351 DUFs detected in 454 varieties across 79 metagenomes, see Supplementary methods online). We visualized these results as network graphs. Vertices (representing DUFs varieties) were connected if their $\rho$ was $\geq$ 0.90. As abundances of 454 DUF varieties were correlated, we enforced a Bonferroni-corrected p-value threshold of $\sim 2.20\times\ 10^5$ (0.01 / 454). We embedded the graph using the Fruchterman-Reingold procedure [Fruchterman and Reingold, 1991]. A minimal spanning tree was visualized after Prim's algorithm [Prim, 1957] to aid visual interpretation (Fig. 6.1). We assigned DUFs to putative functional categories guided by Pfam descriptions and linked literature, color-coding vertices accordingly.

## 6.4 Results

We observed two prominent networks, one dominated by DUFs linked to photosynthetic organisms (Fig. 6.1, II) and another comprised of more diverse members (Fig. 6.1, I). Smaller networks were observed, including one associating DUFs 404 and 407 (Fig. 6.1, III), domains known to co-occur [Goonesekere et al., 2010]. Employing a 'guilty-by-association' approach (Merico et al, 2009), we propagated hypotheses across closely-embedded domains. We thus hypothesized that DUFs in network II (Fig. 6.1), including unassigned DUFs (Fig. 6.1, grey vertices), describe a photobiologically active module. To examine this hypothesis, the taxonomic distributions of the unassigned domains were examined using the Pfam web-interface (`http://pfam.sanger.ac.uk/`). DUFs 1997, 1995, 1830, and 1651 were observed exclusively in phototrophic organisms while DUF2307 appeared to have a higher copy number in Cyanobacteria. DUFs 2214 and 564 showed less striking distributions, however, we speculate they may be involved in pigmentation and C1 metabolism respectively (see Supplementary material online). The larger network (Fig. 6.1, I) was difficult to interpret due to the diverse functions of its members. The centrality of the

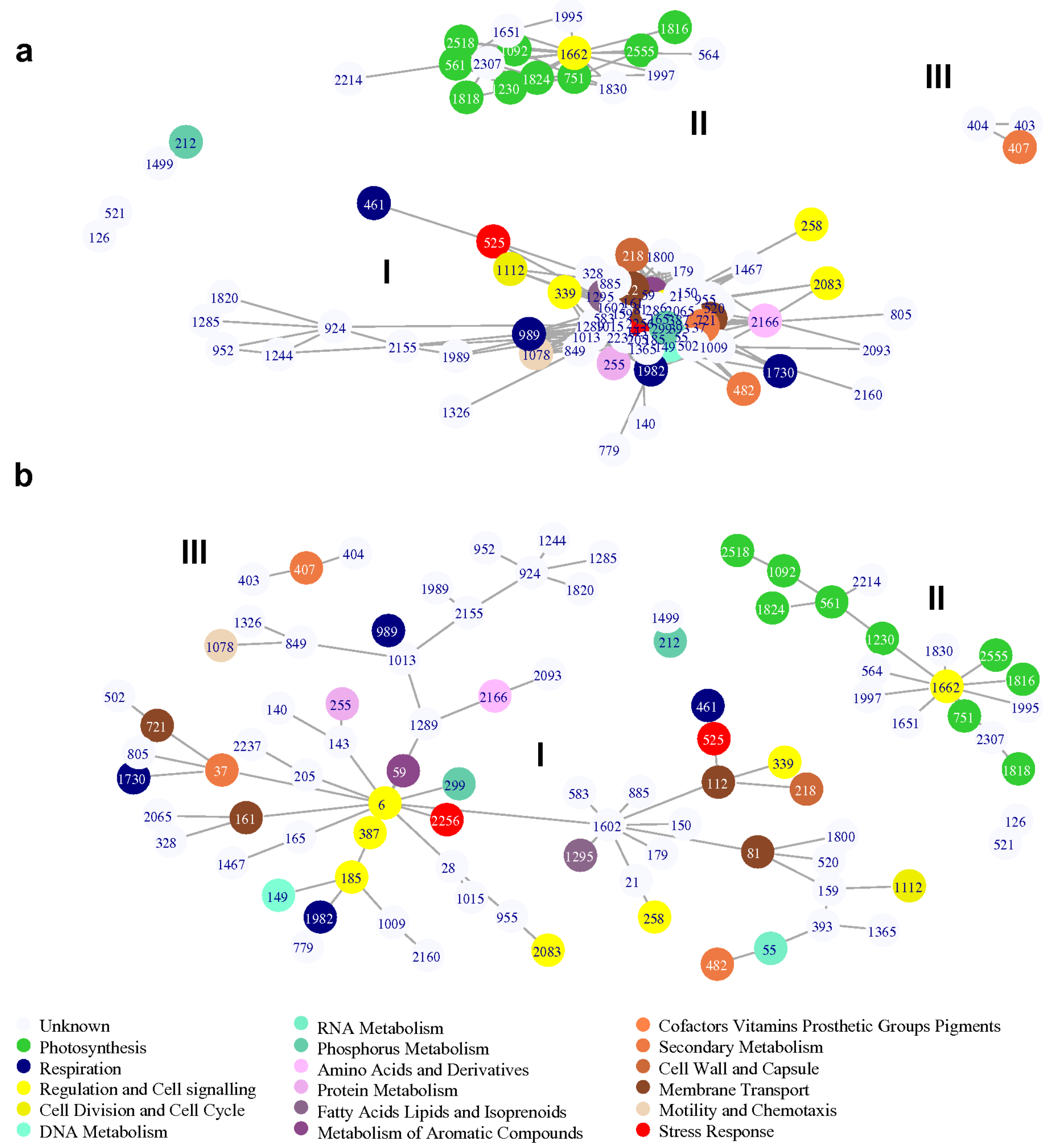


Figure 6.1: Hypothesis generation using network representations of Spearman rank correlations between Domain of Unknown Function (DUF) abundances across GOS metagenomes. Vertices are labelled with the corresponding DUF number (i.e. "59" represents DUF59). Network I includes DUFs with a variety of functions, possibly involved in a response to nutrient input in the marine epipelagic zone. Network II is dominated by DUFs linked to photosynthetic organisms and functions. Network III is composed of three DUFs known to co-occur. a) Fruchterman- Reingold embedded network. Edge lengths are inversely related to correlation strength b) Minimum spanning tree representation of DUF correlations. Only edges describing the shortest path (hence, strongest correlation) between vertices are visualized.

**EamA (PF00892, formerly 'DUF6') domain suggests that its activity may play a role in controlling this functional constellation.**

**Hypotheses generated from covariation across metagenomes may be contextualized with environmental data to enhance interpretation. We**

**employed indirect gradient analysis (see [Ramette, 2007] for a review of multivariate analyses) after [Virtanen et al., 2006] to relate DUF abundances in network II (Fig. 6.1; 17 DUF varieties) to chlorophyll concentrations at appropriate GOS sites (n=56; see Supplementary methods online). We standardized DUF abundances at each site by the median abundance of 22 ‘single-copy domains’ detected at that site.**

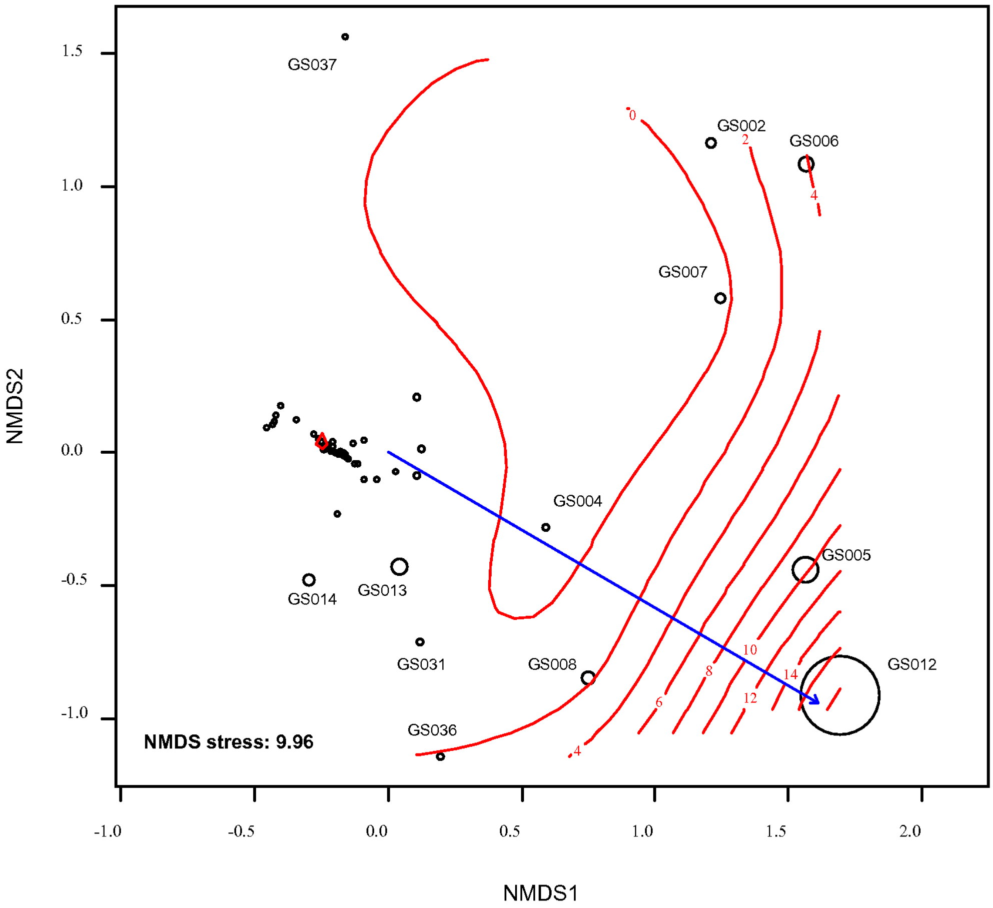


Figure 6.2: Hypothesis contextualization using indirect gradient analysis of GOS sites described by the abundances of DUFs in Network II (ref. Fig. 6.1) and chlorophyll concentration data. Each bubble represents one GOS site and bubble size reflects the in situ chlorophyll concentration measured during the expedition. Bubble positions reflect the Bray-Curtis dissimilarity between sites calculated from the per site abundance profiles of DUFs in Network II. The blue vector describes the linear fit of chlorophyll concentration data to the ordination (Rv $2 \approx 0.52, P(> R) \approx 9.99 \times 10 - 4$). Red isoclines describe a generalized additive model fit of the same chlorophyll data to the ordination (Rs $2 \approx 0.91, p \approx 2.00 \times 10 - 16$) and are labeled with the corresponding chlorophyll concentration (in $4\mu$g chlorophyll per kg seawater). Regions where the chlorophyll isoclines and vector intersect perpendicularly suggest a coherent response gradient of metagenomic DUF content to chlorophyll concentrations.

**We then ordinated sites by non-metric dimensional scaling (NMDS; Fig. 6.2, hollow circles) using Bray-Curtis dissimilarities. Next, we**

performed a least squares, linear fit of chlorophyll data with significance (P(>R)) determined by permutation (n=1000). To explore non-linear relationships between chlorophyll concentrations and the ordination, we visualized generalized additive model (GAM) fits as smoothed, non-parametric isoclines (Fig. 6.2) with significance determined by ANOVA [Wood, 2008]. After Virtanen et al., we interpreted coefficients of determination (R2) as goodness-of-fit measures for linear vectors (Rv2) and non-parametric surfaces (Rs2). Analyses were performed in R (`http://www.r-project.org`). We observed that these DUF abundances moderately, but significantly, structure GOS sites along chlorophyll concentration $(Rv2 \approx 0.52, P(> R) \approx 9.99 \times 10 - 4; Rs2 \approx 0.91, p \approx 2.00 \times 10 - 16)$. An improved, albeit less significant, fit $(Rv2 \approx 0.64, P(> R) \approx 4.00 \times 10 - 3; Rs2 \approx 0.98, p \approx 5.8 \times 10 - 2)$ and a more even resolution of sites may be observed when ordinating geographically localized sample groups such as that along the North American East Coast (GS002, GS004-8, GS012-14; n=9, plot not shown). The GAM surface reveals considerable non-linear effects below chlorophyll concentrations of 2.0 $\mu$g kg-1 seawater, where most sites – particularly from oligotrophic waters – are ordinated. Such effects may rise from the diverse functions, multi-functionality, and the selective interactions between elements in biological systems [Kitano, 2002]. The global coverage of GOS, across numerous ecoregions, may also introduce unexpected variation. Nonetheless, if these chlorophyll measurements are understood as a proxy for phytoplankton abundance, these results tentatively support hypotheses linking the functional community structure described by these DUFs to the abundance of photoreactive plankton. This manner of environmental contextualization may provide useful perspectives on the function of microbial genomic features in their surrounding ecosystems.

## 6.5 Conclusion

Ecogenomic datasets promise to deliver valuable insight into the roles of uncharacterized genes and proteins. The prospects are greater if future ‘omics’ sampling is performed along clear environmental gra-

**dients and accompanied by comprehensive and standardized metadata [Field et al., 2008]. Here, we demonstrated the use of graph theory and numerical ecology to offer new, contextualized hypotheses on uncharacterized targets for community investigation. The restricted composition of the putative photoreactivity module detected by this approach allowed straightforward interpretation. However, estimating globally meaningful false positive and false negative rates is problematic without more ecologically-driven metagenome collections available for comparison. Hence, caution in interpretation, alongside strict detection criteria, is encouraged to reduce association fallacies and hasty generalizations. Subsequently, wet-lab validation of these in silico results is required to establish reliable quality standards for future predictions along a model such as the Computational Bridge to Experiments project [Roberts et al., 2011]; `www.combrex.org`). Navigating the topology of ecogenomic space will be challenging; however, its immense potential in guiding biological enquiry warrants interdisciplinary attention.**

**Supplementary information is available at the ISME Journal's website (`http://www.nature.com/ismej`)**

CHAPTER 7

# SUMMARY AND DISCUSSION

The aim of this thesis was to support biological knowledge generation by the integration of contextual and sequence data. Contributions have been made on the data, information and knowledge level. Different aspects of these contributions are discussed in the following.

## 7.1 Development of MIxS standard-compliant tools

On the data level, two tools to capture and submit contextual data to the public INSDC databases have been developed. Since the standards MIGS and MIMS [Field et al., 2008] and MIMARKS (chapter 4.1) have been introduced to the community only recently (collectively: Minimum Information about any (x) Sequence (MIxS)), it is time to provide the user community with tools that naturally integrate into their daily work. This motivated the development of MetaBar and CDinFusion, which are among the first tools to implement the GSC standards.

Since the applications and workflows in life sciences in general are very heterogenous, it is unlikely that there will be one tool that can be used for all purposes. Therefore, it was clear from the beginning that specialized tools are needed and that not one tool alone can lead to data integration in life sciences.
The focus of MetaBar is to support users to capture contextual data before, during and after sampling campaigns. The barcode concept offers unique identifiers that help to access the data that accumulate

**for a given sample over time along the workflow presented in Figure 1.7 and to keep track of samples.**

**MetaBar received considerable attention after it was published. The BioMed Central Editorial board informed the authors per email, that the article was accessed 2688 times[1] in the first four months after publication. The paper earned the attribute 'highly accessed'. Despite of this fact and even though the tool was successfully applied in several use cases such as the yearly repeating introductory week of the Marine Microbiology (MarMic) Master class on Sylt, in general it has to be stated that the tool has not yet created a wider impact in the life science community that is reflected in a large user community.**

**This can have several reasons. Of course, this additional 'bookkeeping' puts an extra burden to biologists. It can also be that MetaBar will be used more, when the usage of the MIxS standards is promoted more widely. Due to the implementation of a user management with SSL encrypted user authentication, local installations became complicated. User accounts for the MetaBar installation hosted at the MPI Bremen are given to interested users outside the institute. But this service has not significantly been requested. Though it has been tried to implement a straightforward workflow, the integration of various features (e.g. permission concept, exports and submission to INSDC) into MetaBar resulted in a complexity of the tool, that possibly scares off potential users. Though detailed documentation how to use the tool is available, experience collected at the MarMic introductory courses on Sylt, showed that an expert needs a couple of days to teach novices in the usage of MetaBar. Though the contextual data that is captured with MetaBar can be prepared for submission to the INSDC, this was not the primary focus of the tool and the integration of this feature has been done towards the end of the MetaBar development phase.**

**These observations motivated the development of the follow-up tool CDinFusion, with a clear focus on sequence and contextual data submission to the INSDC. It was aimed to avoid many of the negative effects described above, which potentially hamper the impact of such a tool. Since it is required by journals that biologist who analyze sequence data in publications have to submit their data to INSDC, it puts no extra burden to the biologists to prepare these submissions.**

[1] This number is not including access from PubMed Central or other archive sites.

This holds true, if the tool succeeds to do this in a straightforward and easy manner. CDinFusion does not implement a user management. No authentication, logins or passwords are required. Still, CDinFusion - as a web-based tool - can be used in parallel by many users in separate sessions. The workflow of the tool has been reduced to a neccessary minimum. It should qualify to be easy and straightforward. The documentation of the tool has been improved with a seven minute online video tutorial that covers the complete functionality of the tool. The open source code base is readily accessible and can be easily installed. The future will show, if this tool will be accepted by the community.

## 7.2 GSC standards development

The design, development and publication of the MIMARKS standard coincides with the duration of this thesis. After the initial talk at the 6th GSC meeting in October 2008, which started the development of the MIENS standard (later renamed to MIMARKS) under the lead of the MPI in Bremen, the standard was discussed, changed, improved and extended by the GSC community.

For this thesis the development of the MIMARKS standard and the implementation of MetaBar and CDinFusion was a complementary process. The MIGS/MIMS standard [Field et al., 2008] was finished and published during the early implementation phase of MetaBar. The MIMARKS standard was still in development. CDinFusion and the MIMARKS standard were developed in parallel. Changes in the standard led to changes in the implementation of the tool. When requirements for standard improvements became apparent they could directly be communicated. This led to refinements of the standard, while it was still in development. For instance, the need for the GSC web service described in section 4.5, was a result of the implementation of the web forms for the GSC parameters in CDinFusion. In first beta tests of the tool it became apparent that users need detailed descriptions about the GSC parameters. The web service that can be used to retrieve all parameters and descriptions of the MIxS parameters from the GSC database was implemented at the same time. The web forms in CDinFusion are now dynamically retrieving the parameter names and descriptions using this web service.

The standard development was a transparent and open process. The strength of the community-based standard development process is for sure that many people can contribute. The drawback is the slow pace with which this is done. This is due to frequent reiterations and a very eloborate decision making process. It took almost three years to publish this standard. Because many people and institutions all over the world were involved, this again increases the chance that it will be used by the community.

A confusing fact about the MIxS standards is that they are intended to be minimal, which is also reflected in the names of the standards. The community-based development process, however, lead to the fact, that these standards are comprehensive rather than minimal. It is of course difficult to estimate which parameters are relevant and which are not. Of course not all parameters in the standards are mandatory. Users are confronted with long lists of parameters and need to decide which parameters to use and which not. This process might be overwhelming at first.

For data integration and the contextualization of sequence data, it is a very important step forward that the MIMARKS standard is now published. However, the standards leave plenty of room for improvement. Currently, most of the MIxS parameters are free-text fields. To extract information out of these fields automatically, is not trivial, from a computational point of view [Hirschman et al., 2008]. All possible variations and ambiguities of natural language have to be taken into account to interpret the information in these fields correctly. Correct interpretation cannot be guaranteed. It is recommended to use SI units `http://www.bipm.org/en/CGPM/db/11/12/` for measurements, though this is not strictly enforced. Controlled vocabularies or even ontology terms are, with the exception of the environmental ontology (EnvO) terms, not offered to the users. These are things that should be taken into account for the annually planned updates of the standard.

### Stability versus 'living standards'

It has to be noted that even though the MIxS standards are intended to be 'living standards', it is very important from a programmer's and user's point of view to have stability. The planned annual refinement

**of the standard is in this respect a rather short frequency[2]. In order to have compatibility, consistency and function, stability is needed. Refinements should be done in a way that consistency and compatibility are not hampered.**

**Still, a lot of implementation work needs to be done. Even though the basic mechanisms to store the data are present at the INSDC databases, validation routines have to be implemented. GenBank for example has enforced these routines for MIGS fields after they were available in June 2008, but not yet for MIMARKS.**

**When this is all set, special tools need to be available for all kinds of use case scenarios.**

## Adoption of MIxS standards

**Though the GSC is an open and democratic consortium the GSC members occupy mainly higher hierarchical positions in science. There are not many developers and users among the GSC members. Therefore, it can be observed that the MIxS standards are being propagated to the community 'top-down' rather than 'bottom-up', when professors or group leaders decide, their people should use the MIxS standards. Ultimately, life science journals and the INSDC database providers are in the position to require the deposition of contextual data along with publications. If the contextual data submission becomes a requirement the adoption will follow naturally.**

**If these standards are adopted, it will be worthwhile to refine them 'bottom-up'. For this it would be a favourable approach to observe, what turns out to be best-practices. Tool developers and more importantly users should be asked for improvements. This could be done in surveys, by analyzing the scientific literature and last but not least by analyzing the data that accumulates in the INSDC. If for example a certain parameter is not used over a couple of years, it should be removed from a standard that is intended to be minimal[3]. If best practices are taken into account in the years to come, the chance of a wide impact will greatly increase. The MIxS tracker**

[2] If the standard that defines the winding of screws would be changed annually, it would be very hard for the industry to keep up with the production of standard-compliant screws and for the users it would be even harder to screw things together. The same holds true for software.

[3] Something can be considered to be minimal, if no parts can be removed without loosing content.

http://mixs.gensc.org/report/1 is a good entry point to suggest and discuss improvements 'bottom-up'.

## 7.3 Megx.net: A unified view on the data

The megx.net platform with its on-line Genes Mapserver interface to access georeferenced sequence data, provides users with a unified view on the contextual and sequence data. The platform exemplifies the immediate benefits of data integration. Usage statistics show that especially the Geographic-Basic Local Alignment and Search Tool (BLAST) service, was on average accessed 750 times per month from March 2010 until March 2011. When users get hits for their search sequence, through linking to SILVA and ultimately the INSDC entries, they can access a lot of additional information. Also, a download of the complete list of hits is possible. This shows, that the data integration efforts in this direction are useful to the community. Furthermore, the integration of the environmental parameters such as temperature, nitrate, phosphate, salinity, silicate, dissolved oxygen, oxygen saturation, oxygen utilization, chlorophyll and environmental stability in the Genes Mapservers facilitates sequence data analysis in an environmental context.

The integration of sequence data and environmental parameters by using the (x, y, z, t)-key-data tuple has proven to be very useful. This key-data-tuple truly is a minimal contextual data set that helps to link sequence data to many other data sources. The megx.net platform successfully demonstrates how the interpretability of sequence data is increased through data integration.

The data sources that are currently integrated in megx.net are however only "the tip of the iceberg". The Global Ocean Survey (GOS) data set [Rusch et al., 2007] was the first big marine metagenome. It is reasonable to expect many more. It will require a lot of effort to integrate all these upcoming data sets. Also, the World Ocean Atlas (WOA) 2009 has been released. Megx.net is still using the previous WOA 2005. Further scientific environmental data resources like PANGAEA http://www.pangaea.de/ could be not only linked, but in-

**tegrated. If the MIxS standards are adopted widely, the integration efforts can be significantly reduced. However, if the effort necessary to integrate the data in the megx.net project, can not be done by a handful of people, it will be impossible to keep up with the pace of data accumulation. From this point of view, megx.net represents a proof of principle of a future-oriented approach that should be pursued.**

**The megx.net platform has been classified as being at the interface between information and knowledge. This has been done, because the knowledge extraction itself still requires human input, even though the platform offers the users a lot of useful information[4]. That knowledge can be gained with this approach, has been demonstrated in chapter 6 and is discussed in the following.**

## 7.4 Towards knowledge generation *in silico*

**To exemplify the power of *in silico* knowledge generation, a case study has been conducted (chapter 6). The study presents a computational approach to analyze large-scale metagenomic data sets that are enriched with contextual data in order to generate functional hypotheses. This was done by looking at the co-occurrence of domains of unknown function (DUF) across different habitats and by correlating their occurence with environmental parameters offered by the megx.net platform (chapter 5). If the "x,y,z,t-key-data-tuple" had not been available, this analysis would not have been possible. The *in silico* approach demonstrates a way to quickly process and interpret large metagenomic data sets. The GOS data set was analyzed, which at that point of time included more than 10 million sequence reads from 79 sampling sites, all of which were georeferenced. The Hidden Markov Model (HMM) search returned a number of 473,251 hits. The graphical representation of co-occurring DUF hits in networks provided a quick and intuitive overview about this aspect of the data. Patterns were revealed and allowed interpretation already at this stage. Hypotheses about the function of protein families with previously no known function could be derived through the association to protein families with a puta-**

[4] There are no green buttons, that can be pressed to auto-generate scientific papers that contain explicit knowledge.

**tive function. The interpretation of the graphical results furthermore affirmed these hypotheses, through the application of multivariate statistical methods. It could for instance be shown that the DUFs in the photosynthesis cluster, observed in the network graph, are frequently found in habitats with high chlorophyll values.**

**In a next step, the hypotheses should be tested in wet-lab experiments. For that the genomes of marine organisms should be screened for these DUFs. If cultivated marine organisms contain these DUFs, targeted knock-out experiments may reveal, if the deactivation of these domains of function confirms the hypothesis about their function. If for example a DUF, hypothetically involved in photosynthesis, is deactivated, and the deactivation of this function hampers the ability of the organism to perform this function, finally knowledge about the function has been derived. In a community effort, these hypotheses may be tested on a large-scale. Therefore, this approach demonstrates how to arrive at the knowledge level through the intepretation and application of repeatable methods to contextualized sequence data. Without contextual data this study would not have been possible.**

## 7.5 Getting the most out of the data

**Further strategies that are similar to the approach presented in chapter 6 need to be developed and applied. The aim must be, to make use of the power of computers to process large quantities of data[5] and to make use of the strength of human minds to interpret results and to draw conclusions. This can be an iterative process, where processing strategies are refined after humans interpreted results and developed new ideas how to improve the methods applied. For this, data has to be brought into a computable form. Different data integration strategies as well as the application of different database technologies to store and process the data need to be explored.**

**The enrichment of the primary sequence data with contextual data offers an efficient way and seems to be a good starting point. The formalization of the data through the application of controlled vocab-**

[5] Meant are quantities that exceed the ability of human minds to process data.

**ularies is a logical next step. This could be integrated into the GSC standards. Terms for certain parameters could for example easily be offered to users in user interfaces. This will increase the computability of the data and will lead to results that can be interpreted by humans more quickly.**

**Ontologies offer even more ways to automatically derive new knowledge out the data. However, formalization requires human input. The development of ontologies and the translation of data into ontologies is an elaborate process. It has to be considered, if the effort-to-benefit-ratio justifies the work, which has to be additionally invested.**

**The sheer quantity of the data could make it inevitable, to intensify the use of means to automatically infer knowledge. The requirements for human interpretation could be reduced, by drawing conclusions *in silico*. Solutions to this may be achieved through the application of the rich toolbox of artificial intelligence.**

CHAPTER 8

# CONCLUSION AND OUTLOOK

*"Ignorance more frequently begets confidence than does knowledge: it is those who know little, and not those who know much, who so positively assert that this or that problem will never be solved by science."* (Charles Darwin, The Descent of Man, 1871)

## 8.1 Life Sciences: The transition to a data-driven field of science

As it was discussed, it is an indisputable fact that scientists in all disciplines of life sciences are facing an unprecedented data accumulation. Hence, the field is in a transition phase to a data-driven field of science. Other fields of science like physics or astronomy have long since undergone this transition. The challenges in these fields of science were mastered by the application of efficient data processing strategies and by applying state-of-the-art computer science methods and technologies. In order to gain knowledge, it is inevitable to also apply these strategies to the ever-increasing amount of sequence data in life science.
In this thesis several strategies in this direction have been pursued.

- Early contextual data acquisition and integration: It is necessary to preserve the context of the sequence data early on and to integrate contextual and sequence data as soon as possible to reduce manual input to a minimum and to enhance opportunities to automatically process the data.

- **Standards:** The development, implementation of standards for contextual data have the potential to create a wide impact and to facilitate broad use of the data. This has to be done in a community effort and requires the adoption of the standards.
- **Knowledge:** Finally, efficient strategies to gain knowledge which make use of the available contextual data, need to be developed. An example for that has been given in chapter 6.

## 8.2 Projects on the horizon

The GOS project started the age of large-scale environmental metagenomic data sets.
There are many follow-up projects with specific focuses on the horizon. The TARA ocean cruise (`http://oceans.taraexpeditions.org/`) and the Malaspina project (`http://www.expedicionmalaspina.es/Malaspina/Main.do`) are further exploring the oceans' ecosystems. The human microbiome project (`http://commonfund.nih.gov/hmp/`) is investigating the microbial diversity in humans. The Earth microbiome project (`http://www.earthmicrobiome.org/`) aims to comprehensively characterize the global microbial taxonomic and functional diversity.
There is no doubt that these projects will create vast amounts of sequence data. With the MIxS standards in place, it can be expected that contextual data are recorded and publicly deposited along with these sequence data. Once contextualized, a dense network of data points will be created (schematically depicted as overlapping data clouds about “Organisms“, “Genes“ and the “Environment“ in figure 8.1). The denser this network becomes and the better these data are integrated through the usage of contextual data, the greater will be the scope of analysis possibilities. In data analyses it will become easier to distinguish signal from noise. More and more statistically meaningful signals will be detected in the ever growing integrated data set. In the years to come, the spotlight needs to be on the development of methods and strategies to detect these signals. To draw a hypothetical picture about where the contextual and sequence data integration

might lead us, in the last section of this thesis a ’contextual data utopia’ is described.

## 8.3 A contextual data utopia

Ideally, there will be a multitude of tools and mechanisms that help to capture, process and exchange contextual data in the future.

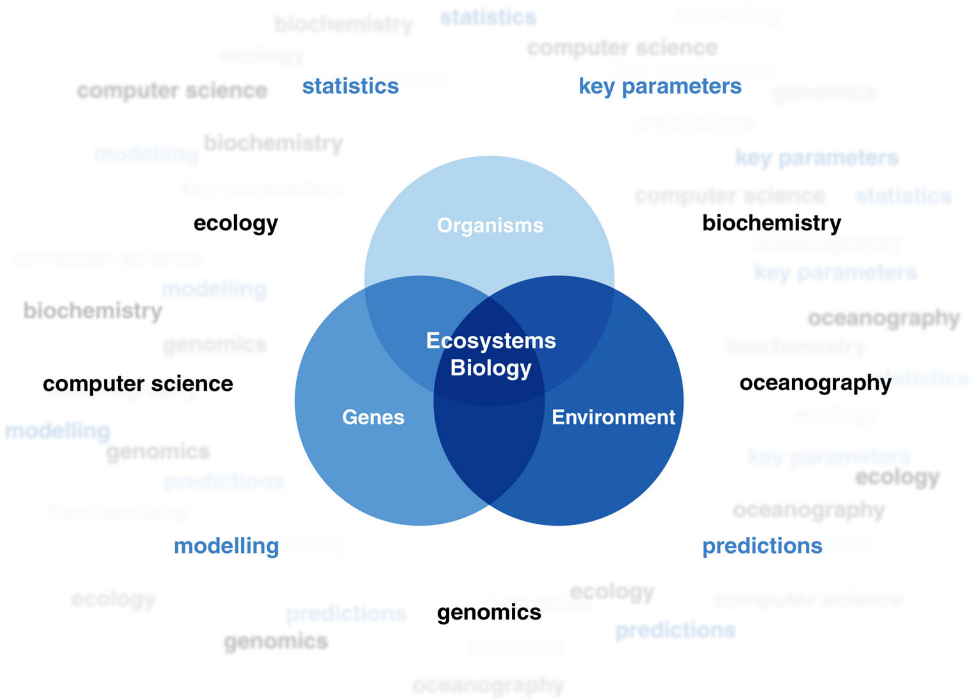


Figure 8.1: A multidisciplinary integrative approach to ecosystems biology. From: [Glöckner and Joint, 2010]

The tools and mechanisms are used by a large community. The quantity and quality of the sequence and contextual data increases dramatically. Knowledge generation is accelerated through data integration in various ways. Standards are so far developed and widely adopted that the data can be processed automatically and knowledge can be derived mainly *in silico*. The methods to process the data are so advanced, that it is easy for humans to interpret the results and to make wise decisions based on them. An ever growing integrated and contextualized sequence data set (compare figure 8.1) combined with a multidisci-

**plinary approach allows ecosystems modelling in great detail and on a global scale. The microbial component in the Earth's biosphere is profoundly understood. Sequencing of environmental data will happen in a high frequency and high throughput manner and realtime monitoring becomes possible. The sequence data is contextualized with high resolution, environmental measurements in realtime, so that the impact of human alterations can be monitored. Global changes can be predicted and steered. The outbreak of epidemics and pandemics can be predicted and prevented. Identifying and locating microorganisms that perform thermodynamically interesting metabolic processes is easy. Studies of microbial genomes will be complemented with knowledge, derived from environmental metagenomic studies. The study of these processes leads to many useful industrial applications that help to solve agricultural and medical problems. New ways to recycle waste and to produce energy will be discovered. Ultimately, breakthroughs in life science help mankind to return to a balanced way of using planet Earth's resources.**

# Bibliography

[Ackoff, 1989] Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16:3–9.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[Amann, 2000] Amann, R. (2000). Who is out there? microbial aspects of biodiversity. *Syst Appl Microbiol*, 23(1):1–8.

[Amann et al., 1995] Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, 59(1):143–169.

[Anonymous, 2009] Anonymous (2009). Metagenomics versus moore's law. *Nature Methods*, 6(9):623–623.

[Arrigo, 2005] Arrigo, K. R. (2005). Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349–355.

[Bateman et al., 2010] Bateman, A., Coggill, P., and Finn, R. D. (2010). Dufs: families in search of function. *Acta Crystallogr Sect F Struct Biol Cryst Commun*, 66(Pt 10):1148–1152.

[Benson et al., 2008] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2008). Genbank. *Nucleic Acids Res*, 36(Database issue):D25–D30.

[Berthel, 1975] Berthel, J. (1975). *Information, in: Handwörterbuch der Betriebswirtschaftslehre*, volume Bd. 2, 4. Auflage. E. Grochla und W. Wittmann.

[Beynon-Davies, 2002] Beynon-Davies, P. (2002). ***Information Systems: an Introduction to Informatics in Organizations.*** Palgrave Macmillan. ISBN 0333963903.

[Bialas et al., 2007] Bialas, J., Greinert, J., Linke, P., and Pfannkuche, O. (2007). Rv sonne fahrtbericht / cruise report so 191 - new vents "puaretanga hou" : Wellington - napier - auckland, 11.01. - 23.03.2007. Cruise Report 10.3289/ifm-geomar_rep_9_2007, Kiel, Germany.

[Binnewies et al., 2006] Binnewies, T., Motro, Y., Hallin, P., Lund, O., Dunn, D., La, T., Hampson, D., Bellgard, M., Wassenaar, T., and Ussery, D. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. ***Funct Integr Genomics.***

[Booth et al., 2007] Booth, T., Gilbert, J., Neufeld, J. D., Ball, J., Thurston, M., Chipman, K., Joint, I., and Field, D. (2007). Handlebar: a flexible, web-based inventory manager for handling barcoded samples. ***Biotechniques***, 42(3):300, 302.

[Breitling et al., 2008] Breitling, R., Gilbert, D., Heiner, M., and Orton, R. (2008). A structured approach for the engineering of biochemical network models, illustrated for signalling pathways. ***Brief Bioinform***, 9(5):404–421.

[Brussaard et al., 2008] Brussaard, C. P. D., Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G., Heldal, M., Kimmance, S. A., Middelboe, M., Nagasaki, K., Paul, J. H., Schroeder, D. C., Suttle, C. A., Vaqué, D., and Wommack, K. E. (2008). Global-scale processes with a nanoscale drive: the role of marine viruses. ***ISME J***, 2(6):575–578.

[Caporaso et al., 2010] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nat Methods*, 7(5):335–336.

[Champney and Kushner, 1976] Champney, W. S. and Kushner, S. R. (1976). A proposal for a uniform nomenclature for the genetics of bacterial protein synthesis. *Mol Gen Genet*, 147(2):145–151.

[Cole et al., 2005] Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M., and Tiedje, J. M. (2005). The ribosomal database project (rdp-ii): sequences and tools for high-throughput rrna analysis. *Nucleic Acids Res*, 33(Database issue):D294–D296.

[Cole et al., 2009] Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., and Tiedje, J. M. (2009). The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic Acids Res*, 37(Database issue):D141–D145.

[Consortium, 2010] Consortium, U. (2010). The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, 38(Database issue):D142–D148.

[Curtis et al., 2002] Curtis, T. P., Sloan, W. T., and Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci U S A*, 99(16):10494–10499.

[de Wit and Bouvier, 2006] de Wit, R. and Bouvier, T. (2006). 'everything is everywhere, but, the environment selects'; what did baas becking and beijerinck really say? *Environ Microbiol*, 8(4):755–758.

[DeLong, 1992] DeLong, E. F. (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci U S A*, 89(12):5685–5689.

[DeLong et al., 2006] DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., Chisholm, S. W., and Karl, D. M. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311(5760):496–503.

[Díez et al., 2001] Díez, B., Pedrós-Alió, C., and Massana, R. (2001). Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rrna gene cloning and sequencing. *Appl Environ Microbiol*, 67(7):2932–2941.

[Etchemendy, 1999] Etchemendy, J. B. . J. (1999). *Language, Proof and Logic*. Seven Bridges Press,U.S.

[Falkowski et al., 1998] Falkowski, Barber, and Smetacek (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, 281(5374):200–207.

[Field et al., 1998] Field, Behrenfeld, Randerson, and Falkowski (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281(5374):237–240.

[Field, 2008] Field, D. (2008). Working together to put molecules on the map. *Nature*, 453(7198):978.

[Field et al., 2008] Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., Vos, P. D., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glöckner, F. O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.-A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., Gil, I. S., Wilson, G., and Wipat, A. (2008). The minimum information about a genome sequence (migs) specification. *Nat Biotechnol*, 26(5):541–547.

[Finn et al., 2008] Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2008). The pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–D288.

[Floridi, 2003] Floridi, L. (2003). Two approaches to the philosophy of information. *Minds Mach.*, 13:459–469.

[Francis et al., 2007] Francis, C. A., Beman, J. M., and Kuypers, M. M. M. (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J*, 1(1):19–27.

[Fruchterman and Reingold, 1991] Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software Pract Ex*, 21(11):1129–1164.

[Fuhrman, 2009] Fuhrman, J. A. (2009). Microbial community structure and its functional implications. *Nature*, 459(7244):193–199.

[Fuhrman et al., 2006] Fuhrman, J. A., Hewson, I., Schwalbach, M. S., Steele, J. A., Brown, M. V., and Naeem, S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci U S A*, 103(35):13104–13109.

[Fuhrman et al., 1992] Fuhrman, J. A., McCallum, K., and Davis, A. A. (1992). Novel major archaebacterial group from marine plankton. *Nature*, 356(6365):148–149.

[Fuhrman et al., 2008] Fuhrman, J. A., Steele, J. A., Hewson, I., Schwalbach, M. S., Brown, M. V., Green, J. L., and Brown, J. H. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci U S A*, 105(22):7774–7778.

[Galperin and Koonin, 2010] Galperin, M. Y. and Koonin, E. V. (20032010). From complete genome sequence to 'complete' understanding? *Trends Biotechnol*, 28(8):398–406.

[Gilbert et al., 2009] Gilbert, J. A., Field, D., Swift, P., Newbold, L., Oliver, A., Smyth, T., Somerfield, P. J., Huse, S., and Joint, I. (2009). The seasonal structure of microbial communities in the western english channel. *Environ Microbiol*, 11(12):3132–3139.

[Gilbert and Maxam, 1973] Gilbert, W. and Maxam, A. (1973). The nucleotide sequence of the lac operator. *Proc Natl Acad Sci U S A*, 70(12):3581–3584.

[Giovannoni et al., 1990] Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., and Field, K. G. (1990). Genetic diversity in sargasso sea bacterioplankton. *Nature*, 345(6270):60–63.

[Glöckner and Joint, 2010] Glöckner, F. O. and Joint, I. (2010). Marine microbial genomics in europe: current status and perspectives. *Microb Biotechnol*, 3(5):523–530.

[Goonesekere et al., 2010] Goonesekere, N. C. W., Shipely, K., and O'Connor, K. (2010). The challenge of annotating protein sequences: The tale of eight domains of unknown function in pfam. *Comput Biol Chem*, 34(3):210–214.

[Green et al., 2008] Green, J. L., Bohannan, B. J. M., and Whitaker, R. J. (2008). Microbial biogeography: from taxonomy to traits. *Science*, 320(5879):1039–1043.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5:199–220.

[Hall, 2007] Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol*, 210(Pt 9):1518–1525.

[Handelsman, 2004] Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, 68(4):669–685.

[Hankeln et al., 2010] Hankeln, W., Buttigieg, P. L., Fink, D., Kottmann, R., Yilmaz, P., and Glöckner, F. O. (2010). Metabar - a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinformatics*, 11:358.

[Hanner, 2009] Hanner, R. (2009). Data standards for barcode records in insdc (bris).

[Henry et al., 2010] Henry, C., Overbeek, R., and Stevens, R. L. (2010). Building the blueprint of life. *Biotechnol J*, 5(7):695–704.

[Henry et al., 2007] Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophys J*, 92(5):1792–1805.

[Hewson and Fuhrman, 2004] Hewson, I. and Fuhrman, J. A. (2004). Richness and diversity of bacterioplankton species along an estuarine gradient in moreton bay, australia. *Appl Environ Microbiol*, 70(6):3425–3433.

[Hirschman et al., 2008] Hirschman, L., Clark, C., Cohen, K. B., Mardis, S., Luciano, J., Kottmann, R., Cole, J., Markowitz, V., Kyrpides, N., Morrison, N., Schriml, L. M., Field, D., and Project, N. (2008). Habitat-lite: a gsc case study based on free text terms for environmental metadata. *OMICS*, 12(2):129–136.

[Hirschman et al., 2002] Hirschman, L., Park, J. C., Tsujii, J., Wong, L., and Wu, C. H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561.

[Howe et al., 2008] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., Pierre, S. S., Twigger, S., White, O., and Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, 455(7209):47–50.

[Huber et al., 2002] Huber, J. A., Butterfield, D. A., and Baross, J. A. (2002). Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. *Appl Environ Microbiol*, 68(4):1585–1594.

[Huber et al., 2007] Huber, J. A., Welch, D. B. M., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A., and Sogin, M. L. (2007). Microbial population structures in the deep marine biosphere. *Science*, 318(5847):97–100.

[Janies et al., 2007] Janies, D., Hill, A. W., Guralnick, R., Habib, F., Waltari, E., and Wheeler, W. C. (2007). Genomic analysis and geographic visualization of the spread of avian influenza (h5n1). *Syst Biol*, 56(2):321–329.

[Jaroszewski et al., 2009] Jaroszewski, L., Li, Z., Krishna, S. S., Bakolitsa, C., Wooley, J., Deacon, A. M., Wilson, I. A., and Godzik, A. (2009). Exploration of uncharted regions of the protein universe. *PLoS Biol*, 7(9):e1000205.

[Kitano, 2002] Kitano, H. (2002). Computational systems biology. *Nature*, 420(6912):206–210.

[Kottmann et al., 2008] Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., Glöckner, F. O., and Consortium, G. S. (2008). A standard migs/mims compliant xml schema: toward the development of the genomic contextual data markup language (gcdml). *OMICS*, 12(2):115–121.

[Kottmann et al., 2010] Kottmann, R., Kostadinov, I., Duhaime, M. B., Buttigieg, P. L., Yilmaz, P., Hankeln, W., Waldmann, J., and Glöckner, F. O. (2010). Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res*, 38(Database issue):D391–D395.

[Krcmar, 2009] Krcmar, H. (2009). *Informationsmanagement*. Springer, Berlin, 5 edition.

[Lenzerini, 2002] Lenzerini, M. (2002). Data integration: a theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '02, pages 233–246, New York, NY, USA. ACM.

[Liolios et al., 2008] Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N. C. (2008). The genomes on line database (gold) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, 36(Database issue):D475–D479.

[Lombardot et al., 2006] Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C., and Glöckner, F. O. (2006). Megx.net–database resources for marine ecological genomics. *Nucleic Acids Res*, 34(Database issue):D390–D393.

[Ludwig et al., 1994] Ludwig, W., Dorn, S., Springer, N., Kirchhof, G., and Schleifer, K. H. (1994). Pcr-based preparation of 23s rrna-targeted group-specific polynucleotide probes. *Appl Environ Microbiol*, 60(9):3236–3244.

[Ludwig and Schleifer, 1994] Ludwig, W. and Schleifer, K. H. (1994). Bacterial phylogeny based on 16s and 23s rrna sequence analysis. *FEMS Microbiol Rev*, 15(2-3):155–173.

[Ludwig and Schleifer, 2005] Ludwig, W. and Schleifer, K. H. (2005). ***Molecular phylogeny of bacteria based on comparative sequence analysis of conserved genes***, pages 70–98. Oxford university press, New York, USA.

[Ludwig et al., 1998] Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., Bachleitner, M., and Schleifer, K. H. (1998). Bacterial phylogeny based on comparative sequence analysis. ***Electrophoresis***, 19(4):554–568. Using Smart Source Parsing.

[Ludwig et al., 2004] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüssmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., and Schleifer, K.-H. (2004). Arb: a software environment for sequence data. ***Nucleic Acids Res***, 32(4):1363–1371.

[López-García et al., 2001] López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001). Unexpected diversity of small eukaryotes in deep-sea antarctic plankton. ***Nature***, 409(6820):603–607.

[Madigan, 2006] Madigan, Michael T & Martinko, J. M. (2006). ***Brock - Biology of Microorganisms***. Pearson / Prentice Hall, 11th edition.

[Mardis, 2008] Mardis, E. R. (2008). Next-generation dna sequencing methods. ***Annu Rev Genomics Hum Genet***, 9:387–402.

[Markowitz et al., 2008] Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M. A., Grechkin, Y., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K., Hugenholtz, P., and Kyrpides, N. C. (2008). Img/m: a data management and analysis system for metagenomes. ***Nucleic Acids Res***, 36(Database issue):D534–D538.

[Martinez et al., 2010] Martinez, A., Tyson, G. W., and Delong, E. F. (2010). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol*, 12(1):222–238.

[Martiny et al., 2006] Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Ovreås, L., Reysenbach, A.-L., Smith, V. H., and Staley, J. T. (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol*, 4(2):102–112.

[Matthew B. Jones, 2006] Matthew B. Jones, Mark P. Schildhauer, O. R. S. B. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Reviews Ecol. Evol. Syst*, 02/06/08:519 – 544.

[Meehan and Baas-Becking, 1927] Meehan, W. J. and Baas-Becking, L. (1927). Iron organisms. *Science*, 66(1697):42.

[Meyer et al., 2003] Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., and Pühler, A. (2003). Gendb–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*, 31(8):2187–2195.

[Meyer et al., 2008] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. A. (2008). The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386.

[Minz et al., 1999] Minz, D., Flax, J. L., Green, S. J., Muyzer, G., Cohen, Y., Wagner, M., Rittmann, B. E., and Stahl, D. A. (1999). Diversity of sulfate-reducing bacteria in oxic and anoxic regions of a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase genes. *Appl Environ Microbiol*, 65(10):4666–4671.

[Moxon and Higgins, 1997] Moxon, E. R. and Higgins, C. F. (1997). E. coli genome sequence. a blueprint for life. *Nature*, 389(6647):120–121.

[Norvig, 2009] Norvig, S. R. . P. (2009). *Artificial Intelligence: A Modern Approach.* Pearson Education, 3rd edition.

[Overmann, 2006] Overmann, J. (2006). *Principles of Enrichment, Isolation, Cultivation and Preservation of Prokaryotes*, volume 1. Springer New York.

[P. Bocij and Hickie, 2005] P. Bocij, D. Chaffey, A. G. and Hickie, S. (2005). *Business Information Systems: Technology, Development and Management for the E-Business.* Financial Times/ Prentice Hall;, 3rd edition.

[Pace, 1997] Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740.

[Parks et al., 2009] Parks, D. H., Porter, M., Churcher, S., Wang, S., Blouin, C., Whalley, J., Brooks, S., and Beiko, R. G. (2009). Gengis: A geospatial information system for genomic data. *Genome Res*, 19(10):1896–1904.

[Pommier et al., 2007] Pommier, T., Canbäck, B., Riemann, L., Boström, K. H., Simu, K., Lundberg, P., Tunlid, A., and Hagström, A. (2007). Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol*, 16(4):867–880.

[Prim, 1957] Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(November 1957):1389–1401.

[Primrose and Twyman, 2003] Primrose, S. B. and Twyman, R. M. (2003). *Principles of Genome Analysis and Genomics.* Blackwell Publishing, 3 edition.

[Pruesse et al., 2007] Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic Acids Res*, 35(21):7188–7196.

[Rahmstorf, 2002] Rahmstorf, S. (2002). Ocean circulation and climate during the past 120,000 years. *Nature*, 419(6903):207–214.

[Ramette, 2007] Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol*, 62(2):142–160.

[Rappé and Giovannoni, 2003] Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annu Rev Microbiol*, 57:369–394.

[Ratnasingham and Hebert, 2007] Ratnasingham, S. and Hebert, P. D. N. (2007). bold: The barcode of life data system (http://www.barcodinglife.org). *Mol Ecol Notes*, 7(3):355–364.

[Richter, 2003] Richter, S. (2003). *Feature-based Programming - Planung, Programmierung, Projekt-Management: Über die Kunst systematisch zu planen und mit Agilität umzusetzen.* ADDISON-WESLEY.

[Roberts, 2004] Roberts, R. J. (2004). Identifying protein function–a call for community action. *PLoS Biol*, 2(3):E42.

[Roberts et al., 2011] Roberts, R. J., Chang, Y.-C., Hu, Z., Rachlin, J. N., Anton, B. P., Pokrzywa, R. M., Choi, H.-P., Faller, L. L., Guleria, J., Housman, G., Klitgord, N., Mazumdar, V., McGettrick, M. G., Osmani, L., Swaminathan, R., Tao, K. R., Letovsky, S., Vitkup, D., Segrè, D., Salzberg, S. L., Delisi, C., Steffen, M., and Kasif, S. (2011). Combrex: a project to accelerate the functional annotation of prokaryotic genomes. *Nucleic Acids Res*, 39(Database issue):D11–D14.

[Rocca-Serra et al., 2010] Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., and Sansone, S.-A. (2010). Isa software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356.

[Rowley, 2007] Rowley, J. (2007). The wisdom hierarchy: representations of the dikw hierarchy. *Journal of Information Science*, 33(2):163–180.

[Rusch et al., 2007] Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M., and Venter, J. C. (2007). The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLoS Biol*, 5(3):e77.

[Salzberg et al., 2007] Salzberg, S. L., Kingsford, C., Cattoli, G., Spiro, D. J., Janies, D. A., Aly, M. M., Brown, I. H., Couacy-Hymann, E., Mia, G. M. D., Dung, D. H., Guercio, A., Joannis, T., Ali, A. S. M., Osmani, A., Padalino, I., Saad, M. D., Savić, V., Sengamalay, N. A., Yingst, S., Zaborsky, J., Zorman-Rojs, O., Ghedin, E., and Capua, I. (2007). Genome analysis linking recent european and african influenza (h5n1) viruses. *Emerg Infect Dis*, 13(5):713–718.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467.

[Schriml et al., 2010] Schriml, L. M., Arze, C., Nadendla, S., Ganapathy, A., Felix, V., Mahurkar, A., Phillippy, K., Gussman, A., Angiuoli, S., Ghedin, E., White, O., and Hall, N. (2010). Gemina, genomic metadata for infectious agents, a geospatial surveillance pathogen database. *Nucleic Acids Res*, 38(Database issue):D754–D764.

[Schulz and Jorgensen, 2001] Schulz, H. N. and Jorgensen, B. B. (2001). Big bacteria. *Annu Rev Microbiol*, 55:105–137.

[Seshadri et al., 2007] Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., and Frazier, M. (2007). Camera: a community resource for metagenomics. *PLoS Biol*, 5(3):e75.

[Sogin et al., 2006] Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J.

(2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103(32):12115–12120.

[Stahl et al., 1984] Stahl, D. A., Lane, D. J., Olsen, G. J., and Pace, N. R. (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal rna sequences. *Science*, 224(4647):409–411.

[Stratton et al., 2009] Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

[Sullivan et al., 2005] Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., and Chisholm, S. W. (2005). Three prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol*, 3(5):e144.

[Taylor et al., 2008] Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P.-A., Bogue, M., Booth, T., Brazma, A., Brinkman, R. R., Clark, A. M., Deutsch, E. W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J. M., Hardy, N. W., Hermjakob, H., Julian, R. K., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Novère, N. L., Leebens-Mack, J., Lewis, S. E., Lord, P., Mallon, A.-M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J. M., Robertson, D. G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R. H., Schober, D., Smith, B., Snape, J., Stoeckert, C. J., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., and Wiemann, S. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. *Nat Biotechnol*, 26(8):889–896.

[Turnbaugh et al., 2007] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164):804–810.

[Tyson et al., 2004] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43.

[van der Staay et al., 2001] van der Staay, S. Y. M., Wachter, R. D., and Vaulot, D. (2001). Oceanic 18s rdna sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, 409(6820):607–610.

[Venter et al., 2004] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74.

[Virtanen et al., 2006] Virtanen, R., Oksanen, J., Oksanen, L., and Razzhivin, V. Y. (2006). Broad-scale vegetation-environment relationships in eurasian high-latitude areas. *J Veg Sci*, 17(4):519–528.

[Vogel et al., 2009] Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., van Elsas, J. D., Bailey, M. J., Nalin, R., and Philippot, L. (2009). Terragenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology*, 7(4):252–252. 10.1038/nrmicro2119.

[Ward et al., 1990] Ward, D. M., Weller, R., and Bateson, M. M. (1990). 16s rrna sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, 345(6270):63–65.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[Whitman et al., 1998] Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A*, 95(12):6578–6583.

[Wieczorek et al., 2004] Wieczorek, J., Guo, Q., and Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18(8):745–767.

[Williamson et al., 2008] Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M., and Venter, J. C. (2008). The sorcerer ii global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One*, 3(1):e1456.

[Wilson, 1991] Wilson, E. (1991). *An introduction to scientific research*. Dover Publications, City.

[Wiltschi and Budisa, 2007] Wiltschi, B. and Budisa, N. (2007). Natural history and experimental evolution of the genetic code. *Appl Microbiol Biotechnol*, 74(4):739–753.

[Winker and Woese, 1991] Winker, S. and Woese, C. R. (1991). A definition of the domains archaea, bacteria and eucarya in terms of small subunit ribosomal rna characteristics. *Syst Appl Microbiol*, 14(4):305–310.

[Woese et al., 1975] Woese, C. R., Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B. J., and Stahl, D. (1975). Conservation of primary structure in 16s ribosomal rna. *Nature*, 254(5495):83–86.

[Wood, 2008] Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *J R Stat Soc Series B Stat Methodol*, 70(3):495–518.

[Yooseph et al., 2007] Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S., Li, H., Mashiyama, S. T., Joachimiak, M. P., van Belle, C., Chandonia, J.-M., Soergel, D. A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B. J., Bafna, V., Friedman, R., Brenner, S. E., Godzik, A., Eisenberg, D., Dixon, J. E., Taylor, S. S., Strausberg, R. L., Frazier, M., and Venter, J. C. (2007). The sorcerer ii global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol*, 5(3):e16.

[Zehr and Kudela, 2011] Zehr, J. P. and Kudela, R. M. (2011). Nitrogen cycle of the open ocean: from genes to ecosystems. *Ann Rev Mar Sci*, 3:197–225.

[Zehr et al., 1998] Zehr, J. P., Mellon, M. T., and Zani, S. (1998). New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (nifh) genes. *Appl Environ Microbiol*, 64(12):5067.

CHAPTER 9

# ACKNOWLEDGEMENTS

First, I would like to thank Prof. Dr. Frank Oliver Glöckner not only for giving me the opportunity to write this thesis in the highly interesting field of bioinformatics, but also for supporting my work and for helping me to master the scientific challenges that a PhD student has to face. Thanks for giving me constructive feedback exceptionally quick. This created a professional and pleasant working atmosphere.

Thanks to Prof. Dr. Peter Baumann and Dr. Wolfgang Ludwig for being supportive members of my PhD thesis and defense committee. Feedback in the thesis committee meetings helped me a lot to keep track.

Before I forget somebody, I want to thank the whole Microbial Genomics Group (MGG). The MGG is really an excellent working environment. I want to thank Norma Wendel for above average commitment in her internship and thesis under my supervision.

Thanks to the whole MarMic faculty. Thanks for the financial support of the Max-Planck-Society and the Jacobs University Bremen. I thank Christane Glöckner and Karl-Heinz Blotevogel for reducing the load of administrative work to a minimum and for being very kind and supportive during my time in the Max Planck Research School of Marine Microbiology Program. Anke Allner, Svenja Frischholz and Peter Tsvetkov at Jacobs University for helping me with administrative issues at Jacobs in an uncomplicated manner.

I would like to thank the MarMic class 2010 and Ivaylo Kostadinov. Thanks for sharing the ups und downs of everyday PhD life! I am grateful that I have found reliable friends here! Thanks to Dennis Fink, Dennis Enning, Mareile Hankeln and Celia Lange for efforts in proofreading of this thesis. And finally, many thanks to my family and my sweetheart Celia for continued support over the past years!