Data Acquisition, Design and Implementation of an Integrative Database for Metagenomics

Renzo Kottmann

April 1, 2005
Contents

1 Introduction ........................................ 4
   1.1 Molecular Microbial Ecology .................. 4
   1.2 MetaFunctions ................................ 8
   1.3 Metagenomics ................................ 11
      1.3.1 Functional screening ................... 12
      1.3.2 Sequence based screening .............. 12
      1.3.3 Metagenome Sequence Assembly ....... 13
      1.3.4 Metagenome Studies and Habitats .... 13
   1.4 Information Extraction from Biological Literature .... 14

2 Material and Methods ............................... 18
   2.1 Platform ...................................... 18
   2.2 OpenGIS ..................................... 18
   2.3 PostgreSQL .................................. 19
   2.4 PostGIS ..................................... 19
   2.5 MapServer ................................... 19
   2.6 GATE: General Architecture for Text Engineering .... 20
   2.7 Databases .................................. 20
   2.8 Corpus collection ........................... 21

3 Results and Discussion ............................ 22
   3.1 Contextual Data Definition .................. 22
   3.2 Description of collected Corpus ............. 24
      3.2.1 Observations on Collected Corpus ...... 24
   3.3 System Outline ................................ 28
   3.4 Database Implementation ..................... 29
      3.4.1 Metagate ................................ 30
      3.4.2 Metamap Database ....................... 30
   3.5 Comparison to other Databases ............... 32
      3.5.1 Marine Bacterioplankton Database (MBD) ... 33
      3.5.2 MicroMar ................................ 33
      3.5.3 PANGAEA ................................ 33
   3.6 The Data Acquisition and Annotation Process ...... 34
      3.6.1 Automatic Database Retrieval Pipeline ... 34
      3.6.2 Automatic Annotation with GATE ........ 35
      3.6.3 Summary: Data Acquisition and Annotation Process ... 35
      3.6.4 Manual Annotation and Evaluation with GATE .... 36

4 Conclusion ........................................ 37
Abstract

The metagenome approach is the culture independent genomic analysis of microbial communities in the environment. It is an emerging new field in Molecular Microbial Ecology based on technical advances in cloning and sequencing protocols. The metagenome approach makes the analysis of large genomic DNA fragments (40-150 kB) from environmental samples possible \cite{Handelsman1998}. It is complementing the SSU rRNA approach in a way that it allows further insight into the genomic content of microbial communities \textit{in situ} beyond 16S rDNA.

In the last 4 years the amount of published metagenome sequences increased exponentially \cite{Streit2004} and a future exponential increase of metagenomic data is anticipated. Not only the vast amount of sequence data, but also the new kind of data requires new storage, processing, and data mining procedures.

Until now there is a clear separation: Metagenome sequence data is deposited in public 'all-purpose' databases and additional environmental data are reported in scientific publications.

To analyse the environmental sequences in their ecological context it is necessary to develop an integrative approach which combines environmental and sequence data.

The “Metagenomes Mapserver” is the first attempt to systematically integrate genomic and metagenomic data into a consistent, curated database including geographic and ecological context information resulting in a Geographic Information System (GIS).

The Metagenomes Mapserver is part of “MetaFunctions” a three year project funded by the European Union (EU). The goal of this project is to elucidate the unknown function of the “conserved hypothetical proteins” by correlating the occurrence of such proteins to habitat specific parameters.

As a first step a database system was designed and implemented in order to combine and curate all available geographic, environmental and metagenome sequence data.

In a second step a total of 77 metagenome paper were acquired and a pipeline for the automatic annotation of the documents was developed.

The result of this work gives the technical basis and required data for the “Metagenomes Mapserver”. This enables MetaFunctions to set up the development of data mining algorithms upon these results. In other words the presented data acquisition, database design and implementation, provides the fundamental technical backbone of the MetaFunctions project.
1 Introduction

Microorganisms can be found everywhere on earth. They live in the air, in terrestrial soils, in oceans, in lakes and rivers, in the sediment in deep subsurface areas, and either as pathogens or as symbionts in eukaryotic multicellular organisms.

They grow under diverse environmental conditions, e.g. oxic and anoxic areas, at extreme low or high temperatures, and with and without light. Some are known to be extremophiles and cannot only survive extreme environmental conditions, but even need these conditions for optimal growth. They thrive in environments where eukaryotic cells can’t survive.

The ability to inhabit almost every place on earth and to gain energy by chemical alteration of their environment makes microorganisms key players in global nutrient cycles such as the sulfur, carbon and nitrogen cycles. Many transformations of sulfur compounds are exclusively carried out by microorganisms. For example sulphate reduction is dominant in anaerobic sediments of the world oceans, which is the biggest sulfur reservoir in the biosphere. [Jorgenson 1982; Widdel and Hansen 1992]. Also anaerobic methane oxidation (AOM) is a globally important process in the carbon cycle that significantly reduces the methane – a greenhouse gas – flux from the ocean to the atmosphere [Reeburgh 1996].

The fact that microorganisms are the only organisms known being able to fix elemental nitrogen (N$_2$) from the atmosphere makes them key player in the nitrogen cycle.

1.1 Molecular Microbial Ecology

Microbial Ecology – a major discipline of Microbiology – is the study of the interaction of microorganisms with their biotic and abiotic environment. The rational for studying the ecology of microorganisms is that these organisms play a key role in the global element cycles and are fundamental for sustaining ecosystems.

The study of microorganisms in their environment is somehow limited by adequate methods [Azam 2001]. The discovery of DNA as the fundamental macromolecule - carrying all the necessary information for maintaining all processes within a cell - and subsequent development of molecular techniques, was not only the foundation of Molecular Biology, but also revolutionized Microbiology.

The advances in molecular techniques gave also rise to the relatively young discipline of Molecular Microbial Ecology. In this discipline molecular methods such as hybridization and sequencing techniques are used and further developed to tackle two main topics:

1. Biodiversity: culture independent identification, and quantification of microorganisms in various habitats. How many of which kind?
2. Microbial activity: What are microorganisms doing in their habitat?

The rRNA Approach. Until the mid 80’s of the last century all studies in Microbiology including those targeting questions related to biodiversity and microbial activity were solely based on culture dependent techniques that is they need pure cultures of single organisms.

Microorganisms isolated using standard cultivation methods are rarely numerically dominant in the communities from which they were obtained (Hugenholtz 2002). They are normally grown on high-nutrient artificial media which favors the rapidly growing ones. Recent estimates suggest that only 1% of the diversity can be assessed by culture-dependent methods (Amann et al. 1995; Curtis et al. 2002).

The first suggestion to clone SSU rDNA (small subunit ribosomal DNA also termed rRNA genes) directly from the environment in 1986 by Olsen et al. (1986) opened a new door for cultivation independent molecular studies of microbial diversity known as the rRNA approach.

With the first cultivation independent studies on microbial diversity in soils (Torsvik et al. 1990), open ocean (Giovannoni et al. 1990), and other habitats it could be shown that microorganisms are the unseen majority (Whitman et al. 1998) and that the microbial diversity is even higher than expected. Numerous others studies on bacterioplankton diversity in open oceans also revealed that the majority of SSU rDNAs belongs to new phylogenetic groups with no close relatives in culture collections (Schmidt et al. 1991; DeLong 1992; Fuhrman et al. 1993; Gonzalez and Moran 1997; Suzuki et al. 1997; Hugenholtz et al. 1998).

The general procedure for the rRNA approach can be summarized in a sequence of experimental steps as follows: From an environmental sample, the total community DNA is extracted then either directly cloned in a host or first selectively amplified, and then cloned. Afterwards the sequence is determined from clones containing the 16S rRNA gene and finally comparative analysis of the retrieved sequences is performed and stored in databases.

The data obtained by the rRNA approach can be taken to design hybridization probes for specific regions on the rRNA genes. Phylogenetic probes labeled with fluorescent dyes are used for Fluorescence in situ hybridization (FISH) (Amann et al. 1995). In addition to diversity studies FISH methods allow to elucidate abundance and structure of microbial communities in situ. The FISH method has become a standard technique in microbial ecology and is widely used today. It also transforms the rRNA approach to a full circle rRNA approach as depicted in Fig. 1.

The wide use of the new methods, and its impact on the knowledge of diversity, abundance and structure of microbial communities in the environment is also reflected by the exponential increase of 16S rRNA sequences in public databases like Ribosomal Database Project II (RDP) and ARB
Figure 1: Flow chart representation of the general full circle rRNA approach. Taken from Amann et al. (1995), courtesy of Prof. Dr. F.O. Glöckner

As on March 28, 2005 the RDP database contains 128,376 bacterial small-subunit rRNA sequences (see Fig. 2 for a growth statistic). The public GenBank database contains at this time 170,808 small-subunit rRNA sequences. In 2003 Rappe and Giovannoni point to the fact that only the major phyla have cultured representatives and the number of 16S rRNA gene sequences from uncultured environmental microbes is roughly three times the number of the cultured counterparts.

Beside the development of culture-independent techniques, pure culture techniques are still fundamental prerequisite for detailed physiological studies and genome sequencing projects.

**Advances in culture techniques.** Advances in isolation techniques are made in the last decade. A high-throughput culturing (HTC) method utilizing FISH for identification was developed (Connon and Giovannoni 2002; Page et al. 2004; Rappe et al. 2002). Another example is the combination of
Figure 2: Diagram showing the exponential increase of 16S rRNA sequences in the RDP database. Each column shows status of RDP after the first update in each year. Compiled by Thierry Lombardot from http://rdp.cme.msu.edu/misc/rel9info.jsp#history

of encapsulation of cells in gel micro-droplets for massively parallel microbial cultivation under low nutrient flux conditions, followed by flow cytometry (Zengler et al. 2002).

In spite of these advances the increase of new microorganisms in culture will not hold pace with the amount of findings by culture-independent techniques (Glöckner and Meyerdierks 2005).

Metagenomics: the next step. The metagenome approach is the extension of the techniques used for 16S rRNA studies (Glöckner and Meyerdierks 2005). With advances in cloning and sequence techniques it became feasible to analyse larger genomic fragments extracted directly from the environment. Thus metagenome libraries are a valuable resource for microbial ecologists for a number of reasons.

- They permit the study of DNA from all microorganisms in a sample, including those that still resist cultivation.
- Each DNA fragment can potentially be placed and interpreted in the context of all other DNA fragments in the same library.
- They contain a vast amount of unexplored DNA, which offers possibilities for the discovery of novel metabolic pathways and enzymatic functions.

So, for the first time it becomes possible to not only determine the diversity of microbial communities, but also gain insights into the metabolic
capabilities of these communities, and especially to get to know the ecological role of the uncultured microorganisms. In this manner the new metagenomics approach is a promising next step. Streit and Schmitz (2004) term metagenomics “the key to the uncultured microbes”. Since the first metagenome study by Schmidt et al. (1991), 77 papers have been published in this domain. They cover a range of aspects varying from methodology, genome reconstruction, discovery of antibiotics, the elucidation of bacterial symbionts to metabolic reconstruction.

Most remarkable is the first report of a proteorhodopsin, which is a light-mediated proton pump, by Beja et al. (2000a) that was found to be abundant in the surface waters all over the oceans. Another interesting finding are archaeal metagenome fragments coding for a set of enzymes normally involved in methanogenesis, together with the extraction of a nickel compound with the same absorption spectrum as the nickel cofactor F430 of the terminal enzyme of methanogenesis from a microbial mat in the Black Sea (Krüger et al. 2003), both underlining the hypothesis of reverse methanogenesis in the anaerobic oxidation of methane.

Both examples show the ability of metagenomics to get new and far reaching insights into the biology of the oceans. The discovery of ubiquitous proteorhodopsin in the ocean surface layer shows the quantitative importance of oceans primary production for the global carbon cycle. And anaerobic methane oxidation is supposed to be the largest sink for this greenhouse gas in marine sediments, and is therefore also significantly shaping the carbon cycle.

Besides this new qualitative gained insights Venter et al. (2004) reported within a single metagenome study based on whole-genome shotgun sequencing 1800 new species, 782 rhodopsin like photoreceptors and more than 1.2 million new proteins. This single study doubled the amount of available protein coding sequences. Most of them are so-called “hypothetical proteins”, proteins with so far unknown function.

1.2 MetaFunctions

For the first time metagenome studies are able to extract and sequence the complete DNA of all organisms in a single habitat and therefore it becomes possible to search for so-called “group specific gene patterns”, which are patterns of genes specific for groups of organism. An artificial example is given in Fig. 3.

Michael Richter was able to find group specific genes of conserved hypothetical proteins for sulphate reducing bacteria in his diploma thesis at the Max Planck Institute of Marine Microbiology. This finding suggests that more such patterns exist for other metabolic groups or even habitats. That could lead to more precise hints for assigning functions to proteins with
unknown functions.

The goal of the “MetaFunctions” project is to correlate metagenome sequence data with ecological data to find group specific gene patterns in order to elucidate the functions of “conserved hypothetical proteins”.

---

Figure 3: Artificial example of a group specific pattern: In Marine bacterium I all four genes coding for a hypothetical protein are present and in one direction of transcription. The second Marine Bacterium has only genes 1, 2, and 4 also in the same direction. Marine fosmid has genes 2, 3, and 4 also in the same direction. Marine fosmid has genes 1, 2, 3, and 4, but gene 2 is in opposite direction. Marine BAC has genes 1, 2, 3, and 4 with gene 1 in opposite direction and a longer stretch of non-coding DNA between gene 1 and 2. The pattern can be summarized as follows: All four genes are in the same order. At least 3 of 4 genes occur in each sequence and a maximum of one gene per sequence is in the opposite direction of transcription.

MetaFunctions is a three year project, funded by the European Union (EU). The full title of this project is “Environmental And Meta-Genomics: A Bioinformatic System to Detect and Assign Functions to Habitat Specific Gene Patterns”. Dr. Thierry Lombardot and Prof. Dr. Frank Oliver Glöckner initiated this project. The official project start will be in the second half of 2005 in collaboration with Technology Transfer Center (TTZ) Bremerhaven, Poznan University of Technology (PUT) in Poland and Global resource Information Database (GRID) Geneva, Switzerland.

In order to discover group specific gene patterns in the metagenome sequences, integration of sequence and environmental data is needed.

Such an integrative approach has to take into account that metagenome and environmental data is dispersed all over the scientific literature and the sequence data is stored separately in public sequence databases. Relevant

---

Microbial Genomics Group at the Max Planck Institute for Marine Microbiology
Head of Microbial Genomics Group at the Max Planck Institute for Marine Microbiology
information like where the sample was taken or which physical and chemical parameters were present during sample collection are only given in papers.

Until now the concept of the public sequence databases don’t give the appropriate possibility to store environmental and metagenome sequence data in a single concept. This is a key problem addressed within this work. A database concept will be developed that allows the integration of environmental and metagenome sequence data in a single concept.

The database implementation, initial data acquisition, and first visualization is the focus of this work. So this work is part of the work packages “acquisition” and “storage and evaluation” within the framework of MetaFunctions (see Fig.4 for an overview of all working packages).

![Figure 4: Work packages as defined by the MetaFunctions project. See legend for details.](image)

Because the amount of metagenome sequences is expected to increase exponentially an Information Extraction (IE) system has to be developed. This system will be able to extract the environmental and geographic data from the biological scientific literature.

The new database has to model the domain of metagenomics in a way that it integrates environmental, geographic and sequence data. The database model has also to anticipate the needs of Information Extraction (IE) and future data mining tasks. Thus a detailed discussion of the metagenomics approach is given in section 1.3 and Information Extraction is discussed in section 1.4.

The result of this work will be the fundamental backbone of the
Metagenomes Mapserver and enables scientists to perform “comparative metagenomics” (Riesenfeld et al. 2004).

1.3 Metagenomics

The first report on an environmental library construction dates back to 1991 (Schmidt et al. 1991).

According to this several synonyms for the metagenomic approach circulate in the literature: environmental DNA libraries (Stein et al. 1996), soil DNA libraries (MacNeil et al. 2001), eDNA libraries (Brady and J 2000), recombinant environmental libraries (Courtois et al. 2003), community genome (Tyson et al. 2004), whole genome shotgun sequencing (Venter et al. 2004), environmental genomics (Delong et al. 1999), ecogenomics (Beja 2004) and other. Until now “metagenomics” is the most common term used (see (Riesenfeld et al. 2004)) and will be used here.

There are several attempts to define what metagenomics is. The first is more technical whereas the second and third are more general in defining the aim of the metagenome approach independent of the technique used:

Definition 1 “Metagenomic libraries are databases of bacterial clones, usually Escherichia coli, carrying DNA fragments that originate from the collective genomes of all organisms present in a particular environment, habitat or assemblage” (Leveau et al. 2004).

Definition 2 “Metagenomics describes the functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample” (Riesenfeld et al. 2004).

Definition 3 “The metagenome approach is the culture independent genomic analysis of microbial communities in the environment” (Riesenfeld et al. 2004).

Finally the emerging metagenome approach does not have a standard or at least most commonly used methodology. The protocols vary with the chosen habitat. The methodology also depends on the aim of the study: searching for new expressed proteins, elucidating diversity (Beja et al. 2002), reconstruct community, metabolic reconstruction, method development (Leveau et al. 2004) or even viral sequencing (Breitbart et al. 2002).

But all metagenomic experiments follow a general scheme:

To uncover the vast amount of information in a metagenomic library the library has to be screened. To reach this objective several strategies are available.
1.3.1 Functional screening

The functional screening is mostly used in the context of biotechnology and is aimed to look for gene-encoded activities. Some examples are the searches for expression of chitinases, lipases, proteinases or esterases (Henne et al. 2000; Rondon et al. 2000), antibiotic production (Wang et al. 2000; MacNeil et al. 2001; Gillespie et al. 2002), biocatalyst activity (Lorenz et al. 2002), metabolic pathways (Henne et al. 1999) and antiporter activity (Majernik et al. 2001).

An absolute requirement for such activity screenings is the expression of foreign DNA in the host strain. A problem arises if the host strain cannot provide the proper transcription factors for a particular gene, as it will not be able to express the corresponding activity. Proven solutions to this problem are the use of alternative host strains (Wang et al. 2000) or expression vectors that stimulate the transcription of cloned genes independent of their native promoter (Henne et al. 1999, 2000).

1.3.2 Sequence based screening

A second type of screening approach is targeted against elucidation of the DNA sequence. This is accomplished by screening for clones that carry DNA inserts with a phylogenetic marker such as the 16S rRNA gene. The
most common method for 16S rDNA screening is the polymerase chain reaction (PCR) in combination with primers that are specific for individual species, genera or higher taxa (Vergin et al. 1998; Beja et al. 2000a; b; Quaiser et al. 2002; Liles et al. 2003; Lopez-Garcia et al. 2004).

A bonus feature of phylogenetic screening is the unique opportunity to mine the DNA flanking the phylogenetic marker gene for functions that could disclose details on the physiology or ecology of the organism from which the DNA originated (Stein et al. 1996; Beja et al. 2000a, 2002; Quaiser et al. 2002; Liles et al. 2003; Lopez-Garcia et al. 2004). One of the most striking illustrations of this is the discovery of a bacteriorhodopsin in an uncultured marine bacterium, suggesting a lifestyle that depends on energy from sunlight (Beja et al. 2000a).

PCR based screening for other functional genes has the drawback that screening is only possible for “novel variants of known functions classes of proteins” (Daniel 2004).

1.3.3 Metagenome Sequence Assembly

An yet unsolved problem is the assembly of the sequence fragments contained in a metagenome library. This question is especially raised by the ‘while genome shotgun’ approaches like Venter et al. (2004); Tyson et al. (2004); Hallam et al. (2004). To be more precisely, it is still not perfectly solved in single genome projects. The assembly task in metagenomics is complicated, because there is the additional question to which species a particular genome fragment belongs to.

1.3.4 Metagenome Studies and Habitats

The MetaFunctions project aims to correlate environmental parameters and other contextual data like sampling date and geographic position of the sampling site with metagenome sequence data. Hence, it is important to monitor the most significant chemical and physical gradients for microbial life in an environment.

First of all it is necessary to clarify the term habitat:

Definition 4. “The habitat is defined by distinct physical and chemical conditions” (Stolp 1996).

The term habitat and environment are often used synonymously. The microbial community, as the set of microorganisms living together, and the habitat together represent an ecosystem.

Although many microbes have a global distribution, not every species is found in every habitat. This will depend on the complex biotic and abiotic conditions of an environment.
For extreme environments the main factors are e.g. temperature resistance in hot springs or salt tolerance in salt lakes.

This work focuses on the ocean as a marine environment which itself harbors many distinct habitats like sediment, surface waters, deep sea, vents, and seeps.

**Marine Environment**

**Open Ocean** 71% of the earth’s surface is covered by oceans, therefore ocean water is the dominant environment. The open ocean shows more stable environmental conditions as other environments like the soil do. In large parts of the oceans the temperature is never higher than 4°C. In general nutrients are comparatively even distributed.

Sea water has higher concentrations of sodium, chloride, sulphate, and magnesium than fresh waters like lake waters \[s\] \[\text{Stolp} \ (1996)\].

The dissolved organic matter (DOM) content in ocean waters is one of the largest reservoirs of organic carbon \[s\] \[\text{Stolp} \ (1996)\].

Oceans are also the largest anaerobic ecosystems on earth. The Black Sea is anoxic starting from a range of depth of 150 m to 2000 m. One of the most important ecological parameter in the water column is light.

**Sediments** There is only slow diffusion of oxygen from the atmosphere into the water column. Most of the oxygen is respired in the water column. Hence, only a fraction of the dissolved oxygen reaches the ocean floor. As a consequence sediments have anoxic conditions often starting in few millimeter depth.

**Deep Sea** The average ocean depth is 3,800 meter which makes the deep sea the largest habitat in the ocean. The deep sea is characterized by high pressures, which is on average 400 atm, absence of light and temperatures between 2°C - 3°C. Because of the low temperature this habitat does not only need barotolerant and/or barophilic but also psychrotolerant and/or psychrophilic microorganisms.

### 1.4 Information Extraction from Biological Literature

The amount of papers published in the domain of metagenomes is expected to increase exponentially and therefore an Information Extraction (IE) System is needed. Such a system does not try to analyze texts in its whole complexity like natural language processing approaches. Instead the task is to find the parts of a text which contain the relevant information and at the same time 'drop' irrelevant passages. In general the definition of what is relevant is given in specified domain specific rules. For instance, one important task is to extract the geographic position of a metagenome sampling
site. In order to find sentences that most probably deal with the description of the sampling site and its position a naive rule would just look for terms like “(38° 15’ N 45° 78’ E)” as this describes longitude and latitude in the GPS (Global Positioning System) system and further analyse the sentence in which this term occurred to find the site name (e.g. HOT Station) as well.

Typically the information which have to be extracted are modeled with the help of templates. These are sets of attribute value pairs. According to the given example a template would be:

- Longitude/Latitude
- Site Name

The corresponding template instance with the extracted information would be

- Longitude/Latitude = 38° 15’ N 45° 78’ E
- Site Name = HOT Station

The discipline of Information Extraction is quite young and has gained attention in recent years. There is growing demand for Information extraction systems, because of the exploding amount of texts available in the Internet, which needs techniques for the extraction of the relevant information from the flood of data. Remarkably there are two widely recognized conferences that competitively test IE-systems and where the whole IE community meets once a year. One is the “Message Understanding Conference” (MUC) and the other is the “Text REtrieval Conference” (TREC) both hosted by the National Institute of Standards and Technology USA. The latter has a Genomics Track which was in 2004 the track with most participants. Whereas the MUC conference concentrates on text analysis and has no relation to biological literature, TREC’s focus is on Information Retrieval and annotation of gene names in texts.

The relative success of IE in the last years is due to the focus on relevant aspects of the text. The definition needs domain specific knowledge, that is why there exists no robust general IE system. For each particular domain, specific IE systems are developed. Up to now development of IE-systems in biology only focused on genomics and genome annotation (Shatkay 2003).

There is no known IE system in the domain of metagenomics focused on the extraction of environmental data. That makes it necessary to develop an own IE-system from scratch. In this context the aim should be

\[\text{http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html}\]
\[\text{http://trec.nist.gov/}\]
\[\text{http://ir.ohsu.edu/genomics/2004protocol.html}\]
the development of a semi-automated IE-system, because no full-automated system gives 100% correct results. A user interface should enable the user to interact with the Information Extraction process.

A possible architecture is shown in Fig. 6

Figure 6: General system architecture: Information Extraction is applied to literature and a set of attribute value pairs according to a template are generated. A user gets the possibility to view and correct the results before the information is stored for the Metagenomes Mapserver

The steps involved in the development of a semi-automated IE system are:

1. Retrieval of publications in order to construct a corpus
2. Development of a template
3. Manual annotation of publications to construct a test set for evaluation
4. Applying customized IE algorithms to retrieved corpus
5. Development of an interactive user interface

A corpus is a set of texts with additional incorporated linguistic information. The evaluation step needs annotated texts in order to benchmark applied algorithms. A common way of measuring the performance of information extraction algorithms is to determine precision and recall:

\[
\text{Precision} = \frac{|\text{true AVP}|}{|\text{true AVP}| + |\text{false AVP}|} \quad (1)
\]

In this context precision is the number of correct extracted attribute value pairs (true AVP) divided by the number of true AVP added to the number of uncorrect extracted attribute value pairs (false AVP). In other
words, the precision is the fraction of correct found AVPs to all found AVPs. So the higher the precision the more really relevant AVPs are found.

On the other hand, recall is the fraction of correct found AVPs to all extractable AVPs.

\[
\text{Recall} = \frac{|\text{true AVP}|}{|\text{true AVP}| + |\text{not found AVP}|}
\]  (2)
2 Material and Methods

Before the selection of software was made that should aid this work one major general decision was made in advance: the whole MetaFunctions project will use and support Open Source software and Open Standards.

Two main reasons lead to use open source software. First the source code is available: this allows easy customisation and enables the developer to review the code in terms of reliability and security. Second, although many different license models are available, open source software has in common, that the developer is allowed to modify and redistribute the source code and in addition the software is most often free of charge, especially for academic use.

An open standard is created in an open, international, participatory industry process. The open standard is publicly available and free of any license fee. It is thus non-proprietary and owned in common. This makes open standards useful in academic context. Like open software it eases the development process because of the fact that it is non-properietary. An open standard guarantees to high degree that it is interoperable and technology neutral. So the choice of open standard does not determine the software or hardware to use, keeping the freedom to chose other components with own defined criteria. The use of standards in software development is emphasized, because it is essential to ensure reliable, interoperable data and to promote uniformity of data schemes and models in a long term view.

2.1 Platform

The whole system development is based on a clustered Linux environment using the Gentoo distribution\(^9\). The Apache webserver version 2.0.52\(^10\) is installed on this system for the web interface.

2.2 OpenGIS

OpenGIS describes specifications and other products of the Open Geospatial Consortium (OGC) that support transparent access to heterogeneous geodata and geoprocessing resources in a networked environment. The goal of OGC is to provide a comprehensive set of open specifications that enable developers to write interoperating components that provide these capabilities.

\(^8\) For a definition of what Open Source see http://www.opensource.org/docs/definition.php
\(^9\) www.gentoo.org
\(^10\) httpd.apache.org
2.3 PostgreSQL

PostgreSQL is an object-relational database management system (ORDBMS) based on POSTGRES, that was developed at the University of California at Berkeley Computer Science Department. Since 1994 it is further developed as a scalable, SQL compliant database system with different applications including financial data analysis system, a jet engine performance monitoring package, an asteroid tracking database, and a medical information database. In addition PostreSQL is especially used for several geographic information systems.

Several versions of PostgreSQL with different features exist. The newest version 8.0.1 of PostgreSQL is used for the Metagenomes Mapserver because it can be extended with custom defined data types, functions, operators, aggregate functions, index methods and procedural languages. It also supports inheritance of tables. Furthermore several benchmarks suggest that it scales better under high dataload and increasing connections compared to e.g. MySQL.

2.4 PostGIS

For the integration of geographic information the software package “PostGIS” in version 1.0 is used in this work. It adds support for geographic objects to the PostgreSQL server, allowing it to be used as a backend spatial database for geographic information systems (GIS). PostGIS follows the OpenGIS “Simple Features Specification for SQL”.

2.5 MapServer

The MapServer is used as the visualization component providing the map view of the Metagenome Mapserver and the underlying database.

This software is originally developed at the University of Minnesota and continued support is provided through the NASA TerraSIP project.


11 The URL of the PostgreSQL homepage is [http://www.postgresql.org/]
12 There are many ongoing discussions in the database community about measuring database performance. Right now there is no single benchmark available that can compare all available databases with reliable results. So each published benchmark has to be considered carefully. For a comparison of MySQL and PostgreSQL see [http://archives.postgresql.org/pgsql-general/2005-03/thrd2.php#00596]
13 The url of the PostGIS homepage is [http://postgis.refractions.net/home.php]
14 [http://mapserver.gis.umn.edu/]
15 [http://terrasip.gis.umn.edu/]
Although the MapServer is not a full-featured GIS system the MapServer provides a development environment to allow browsing GIS data and creating "geographic image maps".

The MapServer system includes MapScript that allows popular scripting languages such as PHP, Perl, Python, and Java to access the MapServer C API.

2.6 GATE: General Architecture for Text Engineering

GATE\textsuperscript{16} is an infrastructure for developing and deploying software components that process human language. It was developed in the framework of a PhD Thesis \textsuperscript{17}Cunningham 2000 at Sheffield University England. It provides:

1. An architecture for language processing software
2. An Advanced Programming Interface (API)
3. A graphical development environment for the annotation and processing of text documents

GATE provides many third party tools as plugins, it has an Information Extraction component called ANNIE and supports an Ontology-API.

The GATE software is object orientated and written in Java and available as open-source free software under the GNU library licence.

2.7 Databases

The three databases GenBank, PubMed and LinkOut are part of the Entrez retrieval system hosted at National Center for Biotechnology Information (NCBI)\textsuperscript{17}.

**GenBank** The GenBank database is a comprehensive public database of nucleotide sequences \textsuperscript{18}Benson et al. 2004. It is part of the International Nucleotide Sequence Database Collaboration together with the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL). These three organizations exchange data on a daily basis. GenBank grows at an exponential rate\textsuperscript{18}. Each GenBank entry includes a description of the sequence, the scientific name and taxonomy of the source organism and the bibliographic references.

\textsuperscript{16}http://gate.ac.uk
\textsuperscript{17}http://www.ncbi.nlm.nih.gov/
\textsuperscript{18}see growth statistic at http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.htmlfor details
PubMed  The literature database of NCBI is PubMed. It is designed to provide access to citations from biomedical literature. To date it includes over 15 million citations.

LinkOut  LinkOut provides access to full-text articles at journal Web sites and other related Web resources. It also provides access and links to the other Entrez databases. “The goal is to facilitate access to relevant online resources beyond the Entrez system to extend, clarify, or supplement information found in the Entrez databases.”

Beside the web-interface NCBI gives access to its databases with the help of the Entrez Programming Utilities (eUtils). These are a set of seven server-side programs comprising a structured interface to a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature.

2.8 Corpus collection

The corpus was compiled manually. Each full-length paper found was downloaded in the Hypertext Markup Language (HTML) format and Portable Document Format (PDF) if present and stored in the local filesystem with the following naming scheme: “PMIDXXXXXX.html” or “PMIDXXXXXX.pdf” for HTML or PDF, respectively. PMID is shorthand for PubMed identifier and XXXXXXX is the numerical PubMed identifier.
3 Results and Discussion

3.1 Contextual Data Definition

In the first place the contextual data, considered to be an important factor in specific habitats, were determined and specified.

The parameters are divided into three groups:

1. Environmental data: these are according to the definition given in 1.3.4 chemical and physical measurements.

2. Sampling related data: Gives qualitative description of the sample location not directly linked to habitat specific parameters or is result of a classification e.g. epipelagic or oligotroph.

3. Technical data: This data set is concerned with parameters describing a clone library construction like what was the smallest filter used or which vector is used for the library.

'Environmental data’ and 'Sampling data’ should also be strictly differentiated from a technical point of view. 'Environmental data' should only consist of numerical results of direct measurements and 'sampling data' of qualitative description of the sampling site which allows for fuzziness.

**Water column:** Pressure is a derived parameter because it is a function of depth. It increases by one atm (atmosphere) per 10 m. So storing the depth is sufficient and pressure can then be calculated.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Environmental Data</th>
<th>Unit</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water Depth</td>
<td></td>
<td>m</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>20</td>
<td>°C</td>
<td></td>
</tr>
<tr>
<td>O₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conductivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salinity</td>
<td></td>
<td>psu</td>
<td></td>
</tr>
<tr>
<td>Density</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Light intensity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lon/Lat</td>
<td>34N 43W</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Sampling Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Example</th>
<th>Unit</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Name</td>
<td>BATS</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Location Desc.</td>
<td>Sargasso Sea</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Geographic Region</td>
<td>West Asia</td>
<td>–</td>
<td>continents, countries</td>
</tr>
<tr>
<td>Sampling Date</td>
<td>February 1999</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Technical data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Example</th>
<th>Unit</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Filtration Cut-Off</td>
<td>0.45</td>
<td>μm</td>
<td>–</td>
</tr>
<tr>
<td>Max Filtration Cut-Off</td>
<td>1</td>
<td>μm</td>
<td>–</td>
</tr>
<tr>
<td>Sample Size</td>
<td>–</td>
<td>–</td>
<td>volume or weight</td>
</tr>
<tr>
<td>Cloning Vector</td>
<td>fosmid</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Marker Gene</td>
<td>APS</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Estimated Phylogeny</td>
<td>γ-Proteobacteria</td>
<td>–</td>
<td>taxonomy id</td>
</tr>
<tr>
<td>Publication Reference</td>
<td>John Public</td>
<td>–</td>
<td>PubMedId, doi</td>
</tr>
</tbody>
</table>

### Sediment: The main difference between environmental parameters is made by variations in chemical concentrations considered to be important.

### Environmental Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Example</th>
<th>Unit</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>20</td>
<td>°C</td>
<td>–</td>
</tr>
<tr>
<td>pH</td>
<td>7</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>O₂</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>H₂S</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Ca</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Lon/Lat</td>
<td>34N 43W</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Water Depth</td>
<td>–</td>
<td>m</td>
<td>–</td>
</tr>
<tr>
<td>Sediment Depth</td>
<td>–</td>
<td>cm</td>
<td>–</td>
</tr>
</tbody>
</table>

### Sampling Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Example</th>
<th>Unit</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Name</td>
<td>BATS</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Location Desc.</td>
<td>Sargasso Sea</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Sampling Date</td>
<td>February 1999</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Sediment Type</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
The technical data is the same as for ocean waters (see above).

The technology group of the “International Census of Marine Microbes” (ICoMM) published notes on the matter of sample collection. They present a list of environmental data considered to be important to studies of the marine environment. This list has some commonalities with the data definition of this work. The definitions for environmental data of the water column are nearly the same as the ‘essential contextual data’ elements of the ICoMM list. The environmental parameter are not as circumstantial as ICoMM defined. Two conceptual differences exist, first of all the ICoMM list is much more detailed and broader defined as the definitions of this work. This is due to the fact that ICoMM seeks “biological and contextual data elements that are important to studies of the water column vs the benthos” and this work is focused on metagenome studies. Therefore only technical data related to metagenome library construction is considered whereas ICoMM’s view is on many different experimental procedures.

3.2 Description of collected Corpus

At the beginning of this work, attempts were made to retrieve metagenome paper via search machines. This was not successful even with searches in specialized databases like PubMed or with the use of specialized search engines like the “scholar” search engine from Google. This may be explained by the fact that metagenomics is an emerging field and that there is no common agreement on naming conventions and several synonyms exist (see 1.3).

So the retrieval was done manually, and as to the best of one’s knowledge 77 published papers reporting results of metagenomic studies were found. There are more papers as reported in recent reviews (Riesenfeld et al. 2004; Beja 2004; Daniel 2004). A total of 72 pdfs and 66 html were downloaded, the remaining 5 were not available via the internet.

3.2.1 Observations on Collected Corpus

A manual survey of all retrieved documents was conducted in order to find common properties and to answer the following questions:

- From which habitats were metagenome libraries constructed?
- How many environmental and contextual data are reported?
- Which unit is used for which environmental and technical parameter?
  Are always the same units used?

22 http://icomm.mbl.edu/
23 http://icomm.mbl.edu/index.php?option=com_content&task=view&id=38&Itemid=71
24 Same URL as above
25 http://scholar.google.com
How are sampling locations reported?

Is Information Extraction on abstracts sufficient?

Half (37) of the metagenome studies have been conducted in the marine environment. 27 metagenome libraries are constructed from soil samples and 13 from other environments.

Almost no published paper gives the complete set of the parameters specified in section 3.1.

Same units are used for depth (meter), temperature (degree celsius), and sample volume (liter for water and gram for soil).

Oxygen concentrations are given only by Suzuki et al (2004) and Venter et al (2004) in two different units: mol/L and mol/kg. Salinity is only given by Venter et al (2004) and LeCleir et al (2004), in one case as practical salinity unit (ppu) and in the other as parts per thousand (ppt). Both units are derived in different ways, but they have the same values, so 35 ppt is equal to 35 psu.

Different formats, units and conventions complicate extraction.

In many cases the abstracts seem to contain almost no relevant information. Even the full text is often not sufficient, because many papers give relevant information in supplementary material, e.g. Venter et al (2004); Hallam et al (2004). Extra information is also presented in tables and pictures. In some cases no direct link to supplementary material exists. Especially in the latter case the information is out of the scope of Information Extraction, because it needs non Natural Language Processing (NLP) algorithms and is therefore a research field on its own.

Origin of Sample and Assignment to Libraries

More than 72% have a location description of the sample origin, and only 23% also give geographic coordinates. The other 77% need manual lookup of maps to find out the geographic coordinates. But also the manual lookup is complicated. In some cases (17%) a location description is simply not present (Holmes et al. 2003; Brady et al. 2002, 2004; Brady and Clardy 2003; Healy et al. 1993; Knietsch et al. 2003; MacNeil et al. 2001) or imprecise. An example for the latter is the following text passage: “eDNA used to construct cosmid libraries was isolated directly from samples collected in Ithaca, N.Y., Boston, Mass., and Costa Rica” (Brady et al. 2004). In this case no more information is given.

Often references to other papers are given in the context of location description. Following the references can lead to the correct sampling site as depicted in Fig. 7.

Sometimes these references are ambiguous, a prototypical example is the following statement: “DNA samples for the SSU and LSU rRNA clone libraries were collected as described previously (Massana et al., 1997) from 0
Figure 7: Schematic representation of references made from one publication to the other giving the origin of the sample. Two different publications refer to a third paper, which describes the sampling procedure of a soil metagenome library. The referenced publication itself refers again to another publication giving location description and environmental data.

The BAC library was constructed in Escherichia coli strain DH10B as previously described (29). A 24,546-member metagenomic library of DNA extracted from soil had previously been constructed (32).

PMID10831436

PMID12208279

PMID10831436

Soil was collected from the West Madison Agricultural Research Station in Madison, Wis (4).

PMID8998199

A subsurface (2-10 cm) soil sample was collected in August 1995, from the West Madison Agricultural Research Station (Madison, WI), and stored on ice until processed. The soil type is a Plano silt loam containing 61% sand, 23% silt, and 16% clay, with 1.7% organic matter. The soil pH was 7.0.

and 500 m at a station S297-67-50, at 124.89° W, 35.44° N, 275 km off Moss Landing, CA, USA.” (Suzuki et al. 2001). Does it mean, an old sample from this sampling site is recycled for the actual study, or the protocol for sample collection is the same and a new experiment was conducted? Following the reference “Massana et al., 1997” it turns out that it is definitely not the same location: “Samples were retrieved near the center of the Santa Barbara Channel, approximately 10 miles offshore from the city of Santa Barbara (34° 15’ N, 119° 54’ W; sea floor at 522 m).” (Massana et al. 1997) Even a human reader can not answer the questions, because it stays unclear if the citation is wrong and actually the same location is meant or the reference to Massana et al. (1997) refers just to the sampling procedure. To complicate this example: Suzuki et al. sampled at “0 and 500 m” depth. In contrast Massana et al. (1997) took more samples at depths between 0 and 500 m. At the end there is no positive indication that the protocol could be the same and the reference can not be solved.

A more detailed view on the references made by different metagenome papers reveals more complex relations as depicted in Fig 8.

Assignment of Sequences to Libraries In almost all papers the sequence accession numbers of GenBank are given. A typical sentence is “Nucleotide sequence accession numbers. Nucleotide sequences have been deposited in the GenBank database under accession numbers AY671989,
Figure 8: Shows interrelations of publications made on metagenome libraries. Boxes with “PMID” indicate papers with PubMed IDs. Arrows pointing to a library indicate that the paper reports a new constructed library. Lines with a circle at one end indicate that the paper mentions an old library. “???” are given where the reference could not definitely be clarified.
AY671990 to AY672016, AY672018 to AY672043, and AY675565 to AY675576.”

Two factors make the assignment of the sequences to the respective library complicated. Is the accession number referring to a sequence of library reported in this article. Second, if the accession number refers to a library reported in the article, then it often stays unclear which library the sequence with the corresponding accession number belongs to.

Like the assignment of sampling sites to metagenome libraries, there is also the case that the assignment is simply not possible. For the paper of Venter et al. (2004) only a project accession number is given because a total of 1.811.372 metagenome shotgun fragments were sequenced. Each single fragment deposited in GenBank does neither have the information from which sampling site, nor from which library it originates.

### 3.3 System Outline

The database system that is developed in this thesis has to integrate environmental and sequence data. Therefore several aspects have to be taken into account: Whereas the sequence data is stored in public sequence databases, e.g. GenBank, the environmental data is distributed throughout the scientific literature. This needs the ability to retrieve and merge data from different sources, mainly sequence databases and scientific publications. These data sources are heterogenous under different aspects. The sequence data is available in structured format, the publications are texts published in the semi-structured HTML format or Portable Document Format (PDF). The storage size is uneven: the sequence data ranges in the order of gigabytes as opposed to a few megabytes for the publications.

In addition the developed system has to fulfill several requirements:

1. Scalability
2. Adaptability: The database has to allow changing demands without the need to remodel the whole data structure
3. Reliability
4. Consistency

The requirements altogether fit to the concept of the “Data Warehouse”. This concept is developed in the context of economy and refers to a “Data Warehouse” as a record of an enterprise’s past transactional and operational activities, stored in a database. This can be seen analog to metagenome studies, which are experiments done in the past (operational activities) from

which the data was stored in publications and databases (transactional activities).

A more precise definition is given by Theodoratos and Sellis (1997): “Data Warehousing is an in-advance approach to the integration of data from multiple, possibly very large, distributed, heterogeneous databases and other information sources.”

The integration process of a “Data Warehouse” extracts, translates and filters data from different sources of interest as appropriate. In addition the gained data is merged with relevant information from other sources and stored in a logically centralised repository (Widom 1995).

The advantage in the orientation of the database development towards a “Data Warehouse” is the focus of this concept on the ETL process – that is “Extract, Transform, Load”, as this is the main task in the data acquisition for the Metagenome Mapserver.

Before the concept of the developed pipeline that implements the ETL process is described, the databases and its implementations are discussed in the next two sections.

### 3.4 Database Implementation

In order to implement these requirements, two databases have been designed. One – named “Metagate” – to store and maintain the acquired publications, sequence data and their related so-called metadata. The other database is designed as the GIS-based public database for the Metagenome Mapserver called “MetaMap”.

Two reasons have lead to this decision. First, some data, that has to be stored, is just necessary for the data acquisition task and not needed for public release and/or data mining tasks. Second, the database for data acquisition tasks has other performance needs, than the one for data mining and public access. The former has to have high performance on data update procedures, whereas the latter needs high-performance on query tasks. The third reason is that the partition in two databases gives the possibility to add a layer in-between the two databases for ensuring data quality. This layer can be used for data quality ensurance, which is the process whereby the data in a database is examined in order to discover inconsistencies. This enables data cleaning before public release. Especially the complexity and the sometimes unresolvable relations between publications as shown in Fig. 8 motivated the data quality ensurance layer, because much care has to be taken to avoid redundancies. The prevention of redundancies is the main instrument to avoid inconsistencies in databases.\(^{27}\)

\(^{27}\)Inconsistencies are discussed in virtually every textbook on database design e.g. Heuer and Saake (2000)
3.4.1 Metagate

The Metagate database is used for the storage and maintenance of all data from retrieval, annotation and Information Extraction.

The semantic is focused on the technical aspects of the data acquisition and Information Extraction. Starting with “Article” the model can be explained as follows: One article has many authors, is published in one journal and reports zero or many sequences. One sequence may be also reported in many articles. These information can mainly be gathered with the help of the newly developed data acquisition and annotation pipeline (see 3.6). In addition the table “ExtractedData” holds all data from the Information Extraction task. In principle it implements the template model as described in 1.4 which is a set of attribute value pairs. In the attribute “tempattname” of the table the attribute name of the template is stored and in the attribute “tempvalue” the corresponding value of the template’s attribute is stored. The attribute “libraryname” assigns each attribute value pair to a specific library (see Fig. 9 for details).

3.4.2 Metamap Database

The database model may be best explained by starting with the ‘Article’. (A graphical representation of the Metamap database is given in Fig. 10). One article can refer to many samples which were studied, and also other article may refer to the samples, either to all or to a subset. One or many reported sample(s) originate(s) from one sampling site and many parameters may have been measured while taking a particular sample. Finally many metagenome sequence fragments can be extracted and sequenced from one sample.

This database implements with the help of PostGIS the “OpenGIS Simple Features Specification For SQL” 28. An attribute of type geometry is added to the table “SamplingSite”. This attribute stores the geographic position of the location from which the sample is taken. With the additional tables “spatial_ref_sys” and “geometry_columns”, every OpenGis compliant client Mapserver is able to use the spatial information in order to draw maps and show the location of the sampling sites.

The crucial part in this model are the tables with “Param” as suffix in the names. These tables make use of inheritance. This allows an “Is-A relation” between tables. For instance, the table “WaterColumnParam” is a “Parameter”. Because of this relation “WaterColumnParam” inherits all attributes of “Parameter”, so “WaterColumnParam” has the attributes “waterdepth”, “O₂”, “pH”, “conductivity”, “density”, “lightint” plus “temperature” and “sampleVolume” from “Paramter”. The use of the IS-A relation allows to model and store the parameter according to the habitat from which the sample

Figure 9: UML of MetaGate database
was taken. This model is also memory efficient, because in each table only the parameter are stored, that are necessary to describe the habitat. The alternative would have been to create one large table with all parameters.

The use of inheritance allows an adaptation of the database model to new habitats without the need of changing the whole database structure by just adding another table. In the case that the first metagenome study from the air would be published, the only change that would have to be done is to add a new table named “AirParam” which inherits all attributes from “Parameter”. This table would have an attribute for the height from which the sample was taken.

Because inheritance is also used to model different kind of samples, the database may also be adapted to other needs by adding a new table which inherits from “Sample”. Conceivably are SSU rRNA studies, genome sequence projects, and isolates.

Figure 10: UML of Metamap database

3.5 Comparison to other Databases

No other database, integrating metagenome sequence data with geographic and environmental data, exists. There are two databases georeferencing 16S
rRNA gene sequences and providing map views of the sampling sites: “The Marine Bacterioplanktion Database” and “MicroMar”. In addition there is a plethora of oceanographic and geologic databases, a single database named PANGAEA was chosen for a closer look because this database is hosted by the “Alfred Wegener Institute for Polar and Marine Research” in Bremerhaven, Germany and the “Center for Marine Environmental Sciences” at University of Bremen, Germany and gives access to environmental data.

3.5.1 Marine Bacterioplankton Database (MBD)

The MBD is a marine 16S rRNA gene database with geographic information. They provide an online map tool based on java applets. In addition queries can be done only by GenBank accession numbers.

The last data update was done in 2003. The data is quite outdated regarding the fast increase of 16S rRNA gene sequences (see Fig 2). Furthermore in order to test MBD, 16S rRNA sequences were searched in GenBank with the query “16Smarine 1999” and three accession numbers (AF114695, AJ237765 and AB013259, all marine bacteria) were randomly chosen. Only accession number AF11469 gave a result. The presented information was nothing more than the GenBank entry.

3.5.2 MicroMar

A notable project called MicroMar integrates 16S rRNA sequence data with environmental and geographic data. In this project the data acquistion is done in a purely manual fashion. Data from nearly 600 paper out of thousands of publications was already collected (Giuseppe D’Auria personal communication). Unfortunately the MicroMar web site was not available during 20. March, 2005 and 31. March, 2005 and no publication about this project currently exist.

3.5.3 PANGAEA

PANGAEA – “Publishing Network for Geoscientific & Environmental Data” is an information system and a digital library for archiving, publishing and distributing georeferenced data with special emphasis on environmental, marine and geological basic research. The focus of PANGAEA is clearly on geological data, they have no data on marine species. It has a relational database system with a web interface for initial data search. For complete data access and detailed analysis PANGAEA develops proprietary software.

29 http://halia.hik.se/mbd/start.jsp?page=start
30 As on 31. March, 2005 the statement was: “The last update was done in November 2003” see http://halia.hik.se/mbd/start.jsp?page=browse%20MBD
31 http://egg.umh.es/micromar/
It seems that PANGAEA has a comparable datamodel. A detailed comparison was not possible, because a graphical datamodel is not available.\(^{32}\) Beside datamodel comparison one major difference is that PANGAEA does not conform to OGC standards which might be a drawback for data sharing.

### 3.6 The Data Acquisition and Annotation Process

Because the metagenome sequence data is stored in databases and the environmental data is given in the respective paper, the data acquisition task is twofold.

An intuitive approach would be, to first process the papers with an Information Extraction system in order to extract all relevant information, especially the accession numbers of the reported sequences, then query the sequence databases with the help of the extracted accession numbers and to store all data in the local databases.

In this case the approach is the other way round. Because all metagenome paper are manually downloaded and stored in a local filesystem with the respective PubMed ID (see \(^{28}\)). First the NCBI databases PubMed, LinkOut and GenBank are queried by the PubMed ID for the retrieval of all relevant data and then the papers are automatically annotated with the retrieved data.

#### 3.6.1 Automatic Database Retrieval Pipeline

Almost all metagenome paper are available via PubMed\(^{33}\). This makes it possible to use the “E-Utils” facilities of NCBI\(^{34}\).

A program was developed, which implements an automatic pipeline for retrieval, processing, and storing all relevant data based on NCBI’s E-Utils.

For each PubMed ID data from LinkOut is collected in order to get data about the journal (publisher, internet address of the article etc.) that published the article corresponding to the PMID. In a second step data from PubMed is collected. These data comprise title, author, year of publication etc. and in some cases the GenBank accession number of the sequence reported in the article. For each sequence accession number, the corresponding data from GenBank is collected in a third step. The whole process is shown in Fig. 11.

---

32 the URL [http://www.pangaea.de/Pictures/DataModel.pdf](http://www.pangaea.de/Pictures/DataModel.pdf) was not available
33 The only exception as the date of writing is: Wilkinson et al. (2002) this may be explained by the fact that articles from the journal “Biotechnology Letters” are listed since 2003 and no article from 2002 is present
3.6.2 Automatic Annotation with GATE

One step further the GATE API was used to implement a second program which loads, annotates automatically and stores the articles as GATE compatible documents. The automatic annotation comprises up to now author names, publisher, title of paper, abstract, and sequence accession numbers. This list can be easily expanded to other entities present in the databases like clone label or isolation source.

3.6.3 Summary: Data Acquisition and Annotation Process

To first retrieve all relevant data from the database and then perform a full automated annotation of the documents has the advantage that the entities extracted from the databases and matched in the text are most probable belonging to the correct category. An example is the category 'accession number'. If an accession number is correctly matched in the text, it is automatically clear, that it is a GenBank accession number and nothing else.

Matches are only reliable to a certain degree, which has to be measured according to precision and recall (see 1.4). A manual survey of all metagenome paper showed that matches of 'title of paper' have a precision and recall of nearly 100%.

No matches at all immediatly give valuable information, they indicate errors in either the document text or the database entry.

An example of a hard to find typing error is given in the paper from Gillespie et al. (2002) in the sentence “primers were designed based on sequences in GenBank (accession no. AFO75724)…” The first digit of the accession number is actually an ‘O’ instead of a zero.

A even more striking example was one case where the title was not
matched in the text. A closer look on this case revealed that the document was saved with the wrong PubMed ID, therefore the retrieval revealed the correct title corresponding to the PubMed ID, but the PubMed ID was simply wrong.

These kind of mistakes can be found, if the program reports mismatches between the database entry and the searched text in the document. A more systematical error recovery would need a specified reporting system and could lead to a better curated database.

3.6.4 Manual Annotation and Evaluation with GATE

Each evaluation of an Information Extraction algorithm or system needs a manual annotated corpus. Only the comparison of an automatically annotated corpus to a manual annotated corpus allows evaluation by calculating e.g. precision and recall. Of course the manually annotated corpus has to have no misannotated entity.

There is no known corpus which could be used. This may be due to the fact, that this is the first attempt to apply Information Extraction to metagenome paper. Hence, one task is to annotate a corpus in order to prepare the evaluation. GATE has a Graphical User interface (GUI) for manual annotation and for evaluation aiding in corpus annotation. Still, manual annotation is a time consuming task and has its own pitfalls.\textsuperscript{35}

The finishing of the manual annotation was not possible within the schedule of this thesis, but the automatic annotation of at least some entities will shorten the manual work in future.

\textsuperscript{35} An introduction in corpus-based work is given by \textit{Manning and Schütze} (2003) mainly discussing the issues involved in the setup of a corpus.
4 Conclusion

The technical base for an integrative metagenomic database is prepared. First the desired environmental and contextual parameters for each habitat of interest were defined. Then the database system was developed according to the defined parameters and following the concepts of “Data Warehouses”. In a next step an automatic pipeline for the acquisition of environmental, contextual and metagenome sequence data which are available at the NCBI was developed. The extracted data is used for the initial annotation of the metagenome documents. The last step shortens the manual annotation of the papers and discovers inconsistencies between data stored in the public databases and data presented in a paper.

The developed system is extendable. One little but important finding at the beginning of this work was, that the corpus of metagenome papers is not large enough for performing reliable and statistically significant evaluations. Therefore the idea right from the beginning was to also process 16S rRNA papers, because they are similar to metagenome papers. Indeed a manual survey of several dozens of 16S rRNA paper revealed no significant difference in the way the 16S rRNA libraries are presented. All developed components – the databases and the annotation pipeline – are designed in a way that they are easily extendable to the processing of 16S rRNA papers.

Also genome papers and isolation papers are imaginable.

The Metagenomes Mapserver is unique. The “Metagenomes Mapserver” is the first attempt to systematically integrate genomic and metagenomic data into a consistent, curated database including geographic and ecological context information. In the context of metagenomics there is no known other attempt going in this direction. The MetaFunctions project is also unique in the attempt to establish an automatic data acquisition pipeline and to put efforts in developing an Information Extraction system that will semi-automatically extract data from metagenome publications.

The corpus is complete. The 77 collected documents represent to the best of one’s knowledge the complete set of published metagenome papers.

The contextual data is sparse. Nearly half of the papers don’t report any environmental parameter. The MetaFunctions project also expected the data to be sparse, therefore it is planned to also integrate 'background data’ from oceanographic and geological public databases. That is to not only

36 Actually the first metagenome paper is a 16S rRNA paper \cite{Schmidt1991}. The authors only screened for 16S rRNA genes and no sequencing of any clone fragment is performed.
correlate the sequence data to the contextual data present in the respective paper, but to also get to know what else is known about the sample location and the habitat where the metagenome study was conducted. This leads to the question if the 'background' data will be sufficient to place the metagenome sequence data in an environmental context, because many aspects have to be considered. Questions will be: How near must the geographic position of the 'background data' be to the original sampling site? How important are microenvironments? These will be some of the major questions, the MetaFunctions project has to answer.

Less than half of the published metagenome sequences originate from large-insert fragments. It is assumed that large-insert libraries will have a higher impact on data mining for the search of group specific gene patterns and a subsequent assignment of functions to the set of "conserved hypothetical proteins", because they allow to analyse the gene-neighborhood.

Especially soil libraries are often screened for single genes and only these genes are sequenced. The "whole genome shotgun" (WGS) metagenome studies produce a Hugh amount of sequence data, but the single fragments are too short (1-3 kB) and it is still an open discussion how trustworthy the assemblies of the WGS studies are.

In order to determine the uniqueness of gene patterns it is necessary to store all data from all metagenome libraries from all habitats and to not restrict the search to the habitat of interest. The logic is as simple as important: If searches for uniqueness are restricted two marine environments than the searches are simply not complete, because it is not checked if a pattern of interest does not also occur in e.g. terrestrial environments.

So, sequence size and the geographic distribution of representatives of different habitats will be important constraints. Hence, it is also worthwhile to think about the integration of whole genomes. The "Marine Microbiology Initiative" of the "Gordon and Betty Moore Foundation" is currently funding and coordinating a genome sequence campaign especially of marine microorganisms. The integration of these data could significantly enhance the data pool for the data mining efforts of the MetaFunctions project.

**Prospects.** The technical base is made before the MetaFunctions Projects started. This fact also complicated the database design, because many aspects which had to be included in the design couldn’t be precisely defined in advance. Hence, the main focus was to design a database, which has to be adaptive. A main test of the design quality will be the integration of other geographic referenced databases like PANGAEA and others in order to get 'background' data.

The fact that less then 80 metagenome paper are published yet indicates

37 See [http://www.moore.org/microgenome/micro_list.asp](http://www.moore.org/microgenome/micro_list.asp) for an overview of currently ongoing sequence projects
that MetaFunctions has the chance to accompany the progress in metagenomics right from the beginning.

Clearly, a full implementation of an Information Extraction System is out of the realm of possibility within the time-frame of this thesis. With this in mind the work plan was from the beginning to establish a framework which is open for further development. Especially the finding that an Information Extraction system focused on extracting geographic, environmental and other contextual data can also be applied to the domain of 16S rRNA libraries make it worthwhile to develop this system further. With the development of an automatic annotation pipeline the building of an annotated corpus for evaluation is near completion and developed algorithms can be tested in future.
References


E. F. Delong, C. Schleper, R. Feldman, and R. V. Swanson. Application of Genomics for Understanding the Evolution of Hyperthermophilic and


Gary R. LeCleir, Alison Buchan, and James T. Hollibaugh. Chitinase Gene Sequences Retrieved from Diverse Aquatic Habitats Reveal Environment-


