Max Planck Institute for
Marine Microbiology

School of Engineering and Science

Data Integration for Marine Ecological Genomics

By

Dipl. Inf./ M.Sc. Renzo Kottmann

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

Doctor of Philosophy

Approved Thesis Committee

Prof. Dr. Frank Oliver Glöckner (chair)
Jacobs University and Max Planck Institute for Marine Microbiology

Prof. Dr. Peter Baumann
Jacobs University

Dr. Dawn Field
Centre for Ecology & Hydrology

Dr. Bernhard Fuchs
Max Planck Institute for Marine Microbiology

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 Information derived from published or unpublished scientific work has been cited in the text and listed in the references.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**

Thesis abstract

Many aspects of biological life can only be explained in their ecological context. This was already recognized more than a hundred years ago by Haeckel: “Ecology is the entire science of the relations of an organism to its environment to which we can count in a broader sense all conditions of existence”. Ecosystem changes can be observed directly on the microbial level through the application of molecular methods developed over the last 20 years. These methods have revealed the diversity and functioning of microbial communities and their crucial role in ecosystem functioning.

The advent of metagenomics, defined as “the functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample” (Riesenfeld et al. 2004), allows researchers, for the first time, to perform cultivation-independent studies of the microbial world on the DNA sequence level. Basic questions like “How does the environment influence the gene content?”, and “How does the functional potential encoded therein influence the capacity of a microbial community to interact with the environment?” can now be addressed. Metagenomics can also be used to test the hypothesis that a portion of the genes with no known function are conserved in certain microbial communities and thus may be important for their successful ecological adaptation and survival. However, to achieve a holistic picture of the microbial realm and the complex interactions therein, data and information on basic ecological questions like “Who is out there?” and “What are they doing?” need to be systematically managed. This is a crucial prerequisite to the relation of sequence data to ecological data based on geographic information, an attribute of both datasets.

The results of this thesis can be grouped into a) genomic data standardization and b) software architecture development and implementation of an integrated framework for ecological genomics. The centerpiece of this thesis is the Microbial Ecological Genomics Database (MegDb). In the vicinity of MegDb a set of tools has been developed using ecological geo-referenced DNA sequence data. In summary, MegDb, a new integrated database suitable for ecological genomics based on existing and newly developed standards is now available. The Minimum Information about a Genome Sequence (MIGS) recommendation by the Genomics Standards Consortium (GSC) is an integral part of MegDb serving to increase interoperability. The involvement in the GSC underpins that successful integration projects need to be based on common standards of international scientific communities.

Table of Contents

Introduction

"What we know is a drop. What we don't know is an ocean."
– Sir Isaac Newton

Isaac Newton's metaphorical perspective on human knowledge aptly describes the volume of discovery awaiting those studying the marine ecosystem. It is well known that 70% of Earth's surface is covered by oceans and it is thought that most life-forms are contained therein. However, it is somewhat ironic that in an era where scientists are able to observe oceanic properties like surface temperature and chlorophyll concentration from space, the question what lives in a droplet of water and how it contributes to Earth's biosphere remains largely unanswered.

Marine Microbial Ecology

Besides the one million microbial cells contained in a millilitre of seawater, microorganisms inhabit almost every place on earth in large numbers. Currently, the total number of microbial cells on earth is estimated to be up to 10^{30}, which is 6-8 orders of magnitude higher than the estimated number of stars in the universe (Whitman, Coleman et al. 1998; Agency 2004). Microorganisms grow under all kinds of environmental conditions, e.g. in oxic and anoxic areas, at extremely low or high temperatures, and with or without light. Some are known to be extremophiles, which not only survive under extreme environmental conditions, but even need these conditions for optimal growth. They thrive in environments where eukaryotic cells cannot survive. They live in air, soils, oceans, lakes, rivers, sediments, and in the deep subsurface. They are free living and can live either as pathogens or as symbionts in eukaryotic organisms[1].

In fact, microorganisms provide the foundation of our biosphere. They were the first organisms inhabiting earth, and catalyzed the key step in the evolution of multi-cellular organisms by oxygenating the atmosphere and allowing aerobic respiration to support life.

The ability to inhabit almost every place on earth and the possession of diverse metabolisms,

[1] It was estimated that a single human body is composed of 10^{13} cells, but harbors $10^{14\text{-}15}$ bacterial cells. This is a difference of one to two orders of magnitude.

with which they can gain energy by chemical alteration in a wide range of environments, make microorganisms key players in global cycling of elements such as sulfur, carbon, nitrogen, and phosphorous.

Moreover, in recent years it has become apparent that microorganisms are even catalyzing key steps in the global element cycles. One example is that many transformations of sulfur compounds are exclusively carried out by microorganisms. Sulfate reduction is the dominant process in the anaerobic sediments of the world oceans, which is the largest sulphur reservoir in the biosphere (Jorgensen 1982; Widdel and Hansen 1992).

Another example is anaerobic methane oxidation (AOM), 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). Additionally, as microorganisms are the only ones known to fix atmospheric dinitrogen, they also play a key role in the nitrogen cycle.

The importance and impact of marine microorganisms on the global element cycles was first recognized in the mid 1970s (Pomeroy 1974). Since then many findings shaped the field of marine microbial ecology. Today it is known that marine microorganisms are responsible for more than 50% of the total primary production, which is the transformation of inorganic- to organic carbon (Field, Behrenfeld et al. 1998; Pedrós-Alió 2006) and are responsible for more than 95% of the total respiration (del Giorgio and Duarte 2002; Pedrós-Alió 2006).

Despite the many findings in the field of marine microbiology and ecology, some fundamental questions are still open and under constant debate: What, exactly, is living in a drop of ocean water? What exactly are they doing; which chemical reactions are they performing, at which rate, and under which environmental conditions? How do they interact within communities; and how is the genomic content and regulation of individual cells related to the functioning of the environment they live in?

Microbial Ecological Genomics

Ecology is ``the entire science of the relations of an organism to its environment to which we can count in a broader sense all conditions of existence'' – Ernst Haeckel, 1866

The formulation of the laws of inheritance by Mendel and the discovery of deoxyribonucleic acid (DNA) as life's fundamental macromolecule – carrying all the necessary information for maintaining all processes within a cell – founded the discipline of genetics. The subsequent development of an array of new molecular techniques, especially DNA sequencing techniques and polymerase chain reaction (PCR) was crucial for the foundation of molecular biology and revolutionized biology (Saiki, Gelfand et al. 1988; Vosberg 1989). This is one reason, why the twentieth century has been labelled "The Century of the Gene" in the book of the same title by the scientific writer Keller (2002).

In the twenty first century especially genome DNA sequencing technologies influence environmental science, advance the understanding of microbial ecology (DeLong 2005), and continuously open many new ways to study single organisms and microbial communities on the DNA level. Based on the advances in molecular biology and ecology, the scientific discipline of Microbial Ecological Genomics emerged recently. It can be defined as: "a scientific discipline that studies the structure and function of a genome with the aim of understanding the relationship between the organism and its biotic and abiotic environments" (Straalen and Roelofs 2006).

Ecological studies of single micro-organisms

In contrast to classical genetics, which studies genes one by one, genomics analyses the genome as a unitary whole, under the premise that the function of one gene can only be understood in the context of the other genes in the genome (Straalen and Roelofs 2006). Genome sequencing delivers the genetic inventory which codes for the cellular functions and phenotype of organisms. In ecological genomics the analysis focuses on genes and regulatory mechanisms thought to be important in the interaction with the organism's environment and the adaptation strategies for living in these environments.

Since the sequencing of the first bacterial genome *Haemophilus influenza* (Fleischmann, Adams et al. 1995) in 1995, a total of 911 complete microbial genomes have been sequenced by the end of 2008 (Kyrpides). Most genome projects are either medically or

biotechnologically motivated. The first marine genome sequenced in 1996 is the Archaeaon, *Methanocaldococcus jannaschii* (Bult, White et al. 1996). It a hyperthermophilic organism isolated from a deep-sea hydrothermal. It was sequenced to understand the genetic basis of the evolution and the biological mechanisms that allow it to not only survive but also to thrive in such extreme environments. *Rhodopirellula baltica* is the first marine bacterium whose sequencing was motivated by purely ecological inquiry (Glöckner, Kube et al. 2003). In 2008, 1845 genome projects were ongoing or completed from which more than 230 are microorganisms isolated from the marine environment. This data is a basis for large scale comparative studies of single genomes.

Ecological studies of microbial communities

The structure of a microbial community can be described e.g. by species richness, biomass and all genomes contained therein. The most commonly used approach in determining microbial diversity is based on the sequencing and analysis of the small subunit (SSU) rRNA genes, particularly the 16S rRNA gene in Bacteria and Archaea. The 16S rRNA gene codes for a crucial function in the translation of messenger RNA to peptides. It is an essential gene present in each known microbe and comprises a mixture of highly conserved and more variable regions where nucleotide substitutions do not alter ribosomal function. Because of the high evolutionary pressure assumed to exist on this gene, it is thought not to be subject to frequent lateral gene transfer and therefore fulfils the criteria of a molecular clock (Zuckerkandl and Pauling 1965; Woese and Fox 1977; Madigan, Martinko et al. 2002). Therefore, the 16S rRNA gene is a suitable marker for molecular diversity studies. The first cloning of SSU rDNA directly from the environment in 1986 by Olsen *et al.* (1986) opened a new dimension to cultivation independent molecular studies of microbial diversity by granting the experimental methodology to more accurately describe “Who is out there?” and assess the environmental diversity.

One major finding of 16S rRNA based diversity studies is that microorganisms isolated by the use of standard cultivation methods are rarely numerically dominant in the communities from which they were obtained (Hugenholtz 2002). Recent estimates suggest that only 1% of the diversity can be assessed by culture-dependent methods (Amann, Ludwig et al. 1995; Curtis, Sloan et al. 2002). In addition studies on microbial diversity in soils (Torsvik, Goksoyr et al. 1990), open oceans (Giovannoni, Britschgi et al. 1990), and other habitats have shown that microorganisms are the unseen majority (Whitman, Coleman et al. 1998)

and that the microbial diversity is even higher than expected. Numerous studies on bacterioplankton diversity in open oceans have shown that the majority of SSU rDNAs belongs to new phylogenetic groups with no close relatives in culture collections (Schmidt, DeLong et al. 1991; DeLong 1992; Fuhrman, McCallum et al. 1993; Gonzalez and Moran 1997; Suzuki, Rappe et al. 1997; Hugenholtz, Goebel et al. 1998).

Another means of describing the microbial community structure on the functional level is the metagenomics approach. The advent of metagenomics, defined as “the functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample” (Riesenfeld, Schloss et al. 2004), allows researchers, for the first time, to perform cultivation-independent studies of the microbial world on the genomic level. Basic questions like how does the environment influence the gene content and how does the genomic potential influence the capacity of an organism to interact with the environment can now be addressed. Moreover, metagenomics can be used to test the hypothesis that a substantial portion of the genes without known function are conserved in certain microbial communities and thus may are important for their successful ecological adaptation and survival.

Technically, metagenomics is an extension of the 16S rRNA approach (Glöckner and Meyerdierks 2005). It is based on improvements in DNA extraction techniques and the continuous advancement of sequencing technologies. Metagenomics allows for the first time to

- Reconstruct genomes without culturing
- Elucidate the biochemical interaction of whole microbial communities

In this manner, the new metagenomics approach is a promising new research approach and has gained much interest. Streit *et al.* term metagenomics “the key to the uncultured microbes” (Streit and Schmitz 2004). Tyson and Banfield are more precisely in stating: ``it is now possible to design experiments that integrate genomics, gene expression and proteomics in an environmental context" (Tyson and Banfield 2005).

Microbial Ecological Genomics: a Data Management and Integration Perspective

The volume of data produced by ecological genomic studies, especially metagenomic projects, demands the use of computational techniques for data management and analysis. Even if computer programs are used or developed *de novo* for a given study, most ecological genomic studies are still done based on manual integration. That is, researchers perform their individual analyses based on *ad-hoc,* one time compilation of the information from their studies and from external sources as needed.

Increasingly, ecological genomic studies do not only analyze their own data, but extensively compare it to published data. The availability of data through the internet has made this process possible. However, the sheer amount of data and the vast number and variety of existing databases makes manual integration infeasible. Today's researchers have to face a high number of data resources with markedly different characteristics. Hernandez *et al.* (2004) accurately summarize the database characteristics most pertinent to a user's point of view:

- The highly diverse nature of the data stored,
- The representational heterogeneity of the data,
- The autonomous and web-based character of the sources and the way the data is published and made available to the public,
- The various interfaces and querying capabilities offered by the different sources.

Clearly, the diverse and heterogeneous characteristic of each individual database makes combination and management of different data sources difficult and, at the very least, time consuming. Before the user is able to use any database in a reasonable way, they must devote significant effort and time to understanding the nature of the data, its representation, its availability and the query capabilities they may use.

Furthermore, researchers interested in combining ecology with molecular information at the same time have to merge data from two different worlds: those of environmental sciences and those of sequence data and bioinformatics.

Environmental databases

Environmental databases collect a variety of geo-referenced observations of different bio- and physico-chemical conditions on the earth in time. Hundreds of databases exist worldwide. Among the databases targeting the ocean are namely World Ocean Atlas (WOA), World Ocean Database (WOD), Pangaea, British Oceanographic Data Centre (BODC), and SeaDataNet.

All these databases differ significantly in a) the nature of the data stored, b) the way the data is presented to users, c) the way the data is published and made available to the public.

There is a high heterogeneity in the exchange formats of the environmental databases. All the above named databases use different data formats for delivering data. This overall high heterogeneity can be explained by the fact that standardization of geographic data and processing only recently grow in scope and importance. For example, the National Oceanic and Atmospheric Administration (NOAA) in the USA, which is hosting WOA and WOD, became Principal Member of the Open Geospatial Consortium (OGC) in March 2009. The OGC is an international consortium of more than 370 companies, government agencies, research organizations and universities participating in the development of publicly available geospatial standards for solutions that "geo-enable" the Web.

Several of the OGC standards became ISO standards including Geography Markup Language and Web Map Service. The environmental databases start to adopt the OGC standards. Pangaea for example started to provide Web Map Service access to their geospatial data.

Sequence databases

The data types most interesting for Ecological Genomics are DNA sequence data as well as transcriptomic and protein data.

Hundreds of molecular biology databases (MBD) exist worldwide. They range from very specialized databases on biological entities to more general databases for different types of data[2]. Notably, examples of specialized and manually curated databases include the Protein Data Bank (PDB) for protein structures (Bourne, Westbrook et al. 2004; Berman 2008) and Swiss-Prot for protein sequences (Consortium 2009).

[2] The Database Special Issue published by Nucleic Acids Research gives a yearly overview of available molecular biology databases.

The most important general databases offer archives for all publically available sequence data. These are EMBL[3] (Kulikova, Akhtar et al. 2007; Cochrane, Akhtar et al. 2008) at the European Bioinformatics Institute[4], GenBank[5] (Benson, Karsch-Mizrachi et al. 2009) at the National Institute of Health and the DNA Data Bank of Japan (DDBJ)[6] (Sugawara, Ikeo et al. 2009) at the National Institute of Genetics. All three databases were created in the context of the ``Human Genome Project (HUGO)'' during the 1990's and constitute the International Nucleotide Sequence Database Collaboration[7] (INSDC). The INSDC databases offer various ways of accessing, querying, and representing the comprehensive set of publically available DNA sequences. Moreover, they all provide substantially different additional data sources and data types. However, the INSDC specifies which additional data on "higher order sequence domains and elements within the genome of an organism" should be provided by all INSDC members in the "Feature Table Definition Document" (INSDC 2009). This assigns common rules upon which the exchange of data on a daily bases is made possible. The data items include:

"regions which:

- perform a biological function,
- affect or are the result of the expression of a biological function,
- interact with other molecules,
- affect replication of a sequence,
- affect or are the result of recombination of different sequences,
- are a recognizable repeated unit,
- have secondary or tertiary structure,
- exhibit variation, or have been revised or corrected" (INSDC 2009).

The data items named above represent results and knowledge gained from extensive laboratory and bioinformatics analysis of each respective DNA sequence. This process of assigning additional data items to regions on the DNA is commonly named annotation.

[3] http://www.ebi.ac.uk/embl/

[4] http://www.ebi.ac.uk/

[5] http://www.ncbi.nlm.nih.gov/Genbank/index.html

[6] http://www.ddbj.nig.ac.jp

[7] http://www.insdc.org/index.html

Today, a range of annotation systems exists to manage and combine data from extensive additional analysis based on diverse sets of bioinformatic algorithms and software (Médigue and Moszer 2007) for sequence assembly (Scheibye-Alsing, Hoffmann et al. 2009), gene finding, protein functions assignment (Juncker, Jensen et al. 2009; Rentzsch and Orengo 2009), protein domain prediction (Finn, Tate et al. 2008) , prediction of gene expression and gene regulation accompanied with data from laboratory studies.

The new large-scale metagenomic sequencing projects – which generate 3-2,000 genome equivalents in sequence information per project – bring new challenges and demand further development on software for assembly, gene calling, and annotation. Several new and dedicated database resources have recently emerged to address the current need for large scale metagenomic data management, namely, CAMERA (Seshadri, Kravitz et al. 2007), IMG/M (Markowitz, Ivanova et al. 2008), and the MG-RAST platform (Meyer, Paarmann et al. 2008).

Data Integration for Ecological Genomics

Data integration systems provide single unified views on the combination of data from different sources. To date, no such system exists, which combines sequence and environmental data resources in order to provide a systematic means of analyzing DNA sequences data in the context of its environment. Even conceptually simple data retrieval requests such as “Give me the temperature at the sampling site of my microbial isolate” are far from trivial. A system tailored to such requests would need to provide comprehensive data derived from a multitude of independent single studies and prepare them for integrated ecological analysis.

Different Integration Approaches

Two general categories of integration system exist: 1) materialized integration systems and 2) virtual integration systems. Materialized integration systems gather data from different sources and store them locally in a unified system, whereas virtual integration systems query many remote sources and only store data relevant to a given query locally. In the latter case, the integrated data exists only virtually at the site of the integration system.

The data warehouse architecture clearly belongs to the category of materialized integration systems. Here, data is extracted from different sources, transformed as necessary and loaded

into the local warehouse. In this context, the data gathering process is often named ETL process (for Extract, Transform, Load) and the data warehouse can be seen as a "Unifying Database" specifying one global data model to which all gathered data is transformed to.

The Mediator based architecture was proposed by (Wiederhold 1992) and is a virtual integration approach. A mediator takes a user query on the mediator schema and translates the query into different queries for the different remote schemas. This architecture requires the specification of correspondences between the single mediator schema and the different remote ones, rather than the full transformation of the source data. The correspondences can be achieved in two different ways. With the Global-as-View concept, each query to a mediator schema component triggers nothing else but queries to one or more remote sources. With the Local-as-View concept each remote data source has to provide a single view on its local data which is in accordance to the mediator schema.

The third navigational architecture is – compared to the last two discussed – rather loosely defined. It emphasizes linking all data items between all resources to allow a navigational point-and-click use of the web pages of all resources. The integration is based on a page model, where interconnections between pages, entry points, and content describing metadata are stored.

Which integration approach is preferable depends on several design decision. Firstly, the *Aim of integration* defines the overall goal of the approach. The aims of existing systems vary: on the one hand, portal concepts, which aim to support "an integrated browsing experience for the user" (Hernandez and Kambhampati 2004) are exemplified by SRS (Etzold, Ulyanov et al. 1996; Zdobnov, Lopez et al. 2002), Entrez (Baxevanis 2008) and other integrated portals. On the other hand, some systems aim to integrate data from different sources and add custom data to augment the value of the system while adding new information.

The common *data model* refers to the underlying technology platform chosen to implement the integrative data model. This can vary from simple text models, semi-structured data e.g. in XML, or structured data most often in terms of relational- or object data. The choice on the data model often reflects the data models of the sources. Design decisions also have to be made on the assumed inter-relations of the source data, which can be either complementary, in that two sources deliver data on different aspects of the integrated data model, or overlapping, in that two sources have same or different data on the same aspects of the integrated data model. The latter is indeed a tremendous difficulty in bioinformatics given the hundreds of databases (Hernandez and Kambhampati 2004; Goble and Stevens 2008). Further design decisions include the *user model* and *level of transparency*, both reflect which kind of user and usage of the integration system is expected. Depending on the expertise of the user, access to the data is given in a browsing fashion, which assumes almost no expertise. Some expertise is required in cases where the system offers data retrieval by querying. The querying can be facilitated by either interactive systems or by direct access to the persistent integrated data. The browsing access gives the user also the highest level of transparency where the sources of data might be even totally invisible, whereas the direct access to the system assumes knowledge of the underlying data structure and sources giving the lowest level of transparency.

Unifying aspects for a unified model

As indicated above, there is a great divide between the worlds of environmental and sequence databases. There are many reasons for this separation. One obvious reason is the different nature of data generated and stored. The environmental databases focus on storing observations of the Earth in terms of physical and chemical measurements in space and time. Sequence databases store complex data on aspects of biological entities, mostly derived from molecular analysis.

An integration approach needs one common unifying data model for the integrated system. Commonly, the description of an integrated system on a higher level is given by a domain model. Such a domain model is the conceptualization of a system which describes the various entities involved in that system and their relationships. The goal of a domain model is to document the key concepts and the domain-vocabulary of the system being modeled.

The questions at hand are: can diverse and complex data of environmental and sequence databases be combined in a single domain model? Are there common touch-points (Goble and Stevens 2008) between environmental and molecular sequence data? Is there a common theme or structure under which they can be integrated?

These questions can be positively answered through proof by example, which is to engineer a formalized domain model proving to be capable of integrating molecular sequence data as well as environmental data.

Environmental data is in most cases a series of measurements in time and space. A simplified schema (using the definition and notation by Garcia-Molina (2002 p. 62)) for integration is

```
Measurement(name, value, unit, x, y, z, time)
```

The attribute are defined as follows: `name` is the name of measurement e.g. temperature; `value` is the measured numerical value e.g. 22; `unit` is the unit of measurement e.g. Kelvin (k); `x` and `y` are the geographic coordinates in a two-dimensional space e.g. longitude and latitude; `z` is the altitude of the measurement location and `time` is the time when the measurement was conducted.

The key point for combining this model with molecular sequence data is to understand that each sample from which sequence data are generated is, by necessity, derived from a sampling event in space and time. This can be modelled as a schema

```
Sample(label, x, y, z, time),
```

where `label` is a name of the sample; `x`, `y`, `z` are again the coordinates in space; and `time` is the time of sample collection. Therefore, the common anchor or the touch-point between these two schemas is geo-referencing, which is the proper recording of the geographic location and time when collecting the sample. Indeed, making the geo-referenced biological sample a key concept of a domain model is a promising approach to enable an integration strategy for environmental and molecular sequence data in ecological genomics.

Experimental Origin of Molecular Sequence Data

Independent of their goals, all biological studies start with one or more samples collected somewhere at sometime. The aims, however, determine how biological samples are treated and what data emerges from them. In the context of ecological genomics the study's scope may be to obtain a single genome of environmental importance, to study the biodiversity or the presence and diversity of key genes, or to obtain the whole genetic markup of a microbial community. The generation of new data and specimens is a shared consequence of virtually all biological experiments. In order to obtain a better overview of how a unifying domain model could be designed, the important specimens for ecological genomic studies are surveyed below. There are two key aspects of the survey: first what kind of data is produced from each treatment and, secondly, where it is deposited and published.

The field sample

Sampling scenarios and procedures acquire material suitable for analysis. All samples can be geo-referenced with standard GPS devices. In addition to the exact time of sampling all additional measurements accompanying the sample collection like temperature, weather conditions etc can be recorded. The additional *in situ* data is as diverse as sample collection scenarios themselves; whereas gathering geographic information for samples from meadows, oceans, and deserts is a rather straightforward process. A common misunderstanding is that it would not be necessary to record the geographic location of the sampling of a movable object like a whale or a human being. This same misunderstanding exists if the target is a symbiotic microbial community living a host organism, which is not in touch with the outer environment. In the scope of such a particular study the geographic origin might be of negligible importance, however, it might be important for other studies. For example, by placing data in a geographic and environmental context, studies have shown that symbiotic communities are affected by the outer environment of the host (Turnbaugh, Hamady et al. 2009).

Foremost, sampling data is noted in the individual researchers' lab-books. Some laboratories have a Laboratory Information Management System (LIMS) where sampling data can be digitalized and archived. Commonly, the data is published only in scientific literature. Even if more field sampling data was originally recorded, usually only the data necessary to support the results and findings of the study are shown. There are some dedicated digital repositories

for field sampling. In the context of marine sampling data widely recognized resources are e.g. Pangaea in Germany, SISMER in France or the British Oceanographic Data Center (BODC). These resources are usually only used for publication of the data if some policy enforces it. The collected field sample is always the starting material for a plethora of following up experimental treatments.

Pure Cultures of Microorganisms

A laboratory pure culture is defined as *in vitro* growth of a single kind of organism in a vessel. A subsample of the field sample is taken and labor-intensive isolation strategies from dilution series to plating with a plethora of growth media is applied to obtain, ideally, a culture of clonal cells. Clonal cells are microorganisms with identical genomes. The pure culture is the prerequisite for physiological and biochemical characterization and taxonomic classification[8].

The information and data gained is voluminous and heterogeneous. An international standard on classification and nomenclature of microorganisms exists. For acceptance of a new species description, a culture has to be submitted to two independent culture collections and a paper must be published describing the new taxon in detail and proposing a name according to the Linnaean System. Worldwide, there are numerous culture collections which provide browsing access to their catalogue. The data contains the name of the organism, whether it is a type strain, and a citation of the descriptive publication. Sometimes information on culturing technique and growth media as well as the deposited sequence information can be found. Straininfo.net[9] offers a navigational integration of most of the information in culture collections (Van Brabant, Gray et al. 2008).

The rRNA Approach

The general procedure for the rRNA approach in microbial ecology can be summarized as follows: The total community DNA is directly extracted from a sample and either selectively amplified by PCR or directly cloned in a vector. Afterwards clones containing the 16S rRNA gene are sequenced

[8] Indeed, prokaryotic taxonomy is a dynamic research field and especially the species concept of prokaryotes is still under debate (for discussion see Gevers, D., F. M. Cohan, et al. (2005). "Re-evaluating prokaryotic species." Nat Rev Micro **3**(9): 733-739. and Rosello-Mora, R. and R. Amann (2001). "The species concept for prokatyotes." FEMS Microbiol Rev. **25**(1): 39-67.

[9] http://www.straininfo.ugent.be/index.php

Comparative analysis and phylogenetic reconstruction of the retrieved sequences is performed and finally the sequences are submitted to the INSDC databases.

The wide use of the rRNA approach and its impact on acquiring knowledge of diversity, abundance, and structure of microbial communities in the environment is also reflected in the availability of numerous tools for the analysis of and phylogenetic reconstruction based on the 16S rRNA gene. Several specialized databases exist such as Silva (Pruesse, Quast et al. 2007), ARB (Ludwig, Strunk et al. 2004), RDP (Cole, Chai et al. 2007), and Greengenes (DeSantis, Hugenholtz et al. 2006).

Genome Sequencing

Today DNA-sequencing is a key method in molecular biology. In 1975 the inventor of didesoxy-sequencing, Frederick Sanger, started with sequencing five nucleotides per day, nowadays whole genomes are sequenced within days on a routine basis and with a variety of technologies.

However, it is important to note that current approaches to genome sequencing rely on the availability of pure cultures.

The sequencing itself produces only raw sequence data. This can be noted as a series of the letters A, C, G, and T – representing the nucleotides adenine, cytosine, guanine, and thymin respectively. A common goal of genome sequence studies is to enrich the raw sequence data with information from further analysis of e.g. Open Reading Frame (ORF) prediction and gene function prediction which demands integrated annotation systems. Médigue *et al.* (Médigue and Moszer 2007) list 26 software tools, 56 publicly available resource databases and 17 systems and platforms commonly used by the community for the annotation and comparison of bacterial genomes.

A scientific publication of a genome requires the authors to submit the sequences to one of the INSDC databases. Submissions commonly include results of the annotation in one of the many file formats offered and accepted by the INSDC databases. Further conclusions and hypothesis gained from the annotation of genome sequences are then reported in the respective publication.

In the light of the growing number of published genomes several specialized integrated genome databases exist with the aim to facilitate comparative genome analysis (Markowitz 2007). Among others are IMG (Markowitz, Szeto et al. 2008), Genome Reviews (Sterk,

Kulikova et al. 2007), and RefSeq (Wheeler, Church et al. 2004). They all differ significantly in how and what exactly they integrate. Furthermore, they all add different kinds and amounts of data and information to the integrated genomes, but the all store similar data on the publically available genomes. Therefore, the named resources are heterogeneous in their technical realization but have overlapping content.

Metagenomics

Advances in DNA extraction protocols allow direct extraction of DNA from a sample without a selective PCR step. Clone based metagenome approaches either chose the Whole Genome Shotgun (WGS) approach where the total DNA of a sample is cloned in vectors of 2-3 kb insert size or the "large-insert" approach where the DNA is size selected for vectors of 40-150 kb. In both cases the clones are subsequently sequenced. With the development of the next generation sequencing technologies like pyro-sequencing (Ronaghi 2001; Nyren 2007) more metagenome studies omit the cloning step and directly sequence the total DNA extract.

As described for genomic data submission, background information on metagenomic sampling and analysis is given in publications; however, the deposition of the sequence data is more heterogeneous. On the one hand, many sequences, especially from the large-insert metagenome libraries, are submitted to the INSDC databases. But the INSDC databases were not prepared for the tremendous amounts of sequence data generated by large-scale metagenome projects. For example, the Sargasso Sea dataset and the following GOS dataset could not be submitted before publication. Later, these datasets where submitted to the NCBI Trace Archive (Sayers, Barrett et al. 2009), which is not synchronized with the other INSDC members.

The initial submission problem is not the only challenge the metagenome sequence datasets pose to the bioinformatics world. The sheer data volume drives existing tools and systems to their limits, and the nature and variety of the information which is gained from these sequence sets puts additional pressure on the research community. The genomic sequences harbor a closed set of genes, which is the complete genetic complement of a single clearly identified organism. Metagenome datasets contain a huge number of sequence fragments in most cases with not more than a single gene from a diverse range of un-indentified and often also unknown organisms. Projects like CAMERA (Seshadri, Kravitz et al. 2007), IMG/M (Markowitz, Ivanova et al. 2008), and MG-RAST (Meyer, Paarmann et al. 2008) have recently emerged to tackle the current needs for large scale metagenomic data management

and analysis.

A Domain Model for Ecological Genomics

The discussion on the experimental origin of molecular sequence data might already indicate that the samples produced during the workflow of a study are the key entities of a common unifying domain model. **Error! Reference source not found.** summarizes the key samples which are produced during the course of typical studies in ecological genomics.

The focus is on modeling real world samples. Data which is, from an ecological perspective, important for analysis and interpretation is then related to the sample from which it was derived or to which it can be meaningfully connected with. For example, an ocean water sample is collected October 15th, 1975 in the North Sea at Helgoland Roads (54°11.3'N, 7°54.0'E) for a metagenomic study. From the perspective of ecological analysis and interpretation of the DNA sequences the important data are the sampled material "ocean water", the calendar datum "15.10.1975", place name "Helgoland Roads, North Sea" and the geographic position "54°11.3'N, 7°54.0'E". While the term "ocean water" is categorical, explicitly derived from the sampling event, all other data are not derived from the sample, but can meaningfully be connected to its entity. Many more data such as the ambient water temperature during the sampling event can be derived or attached to the field sample entity. The criterion for selecting certain data items and relating them to the samples is the usefulness for ecological analysis. That is the question if a data item once recorded helps or is even necessary in later analytical ecological analysis. For example: On the one hand the usefulness of recording the geographic location and time of sampling is obviously necessary from an ecological perspective to be able to analyze the experimental data in a spatial-temporal context and allows linking the individual study to other existing data. On the other hand, data like who conducted the sampling, with which vessel from which stock with which

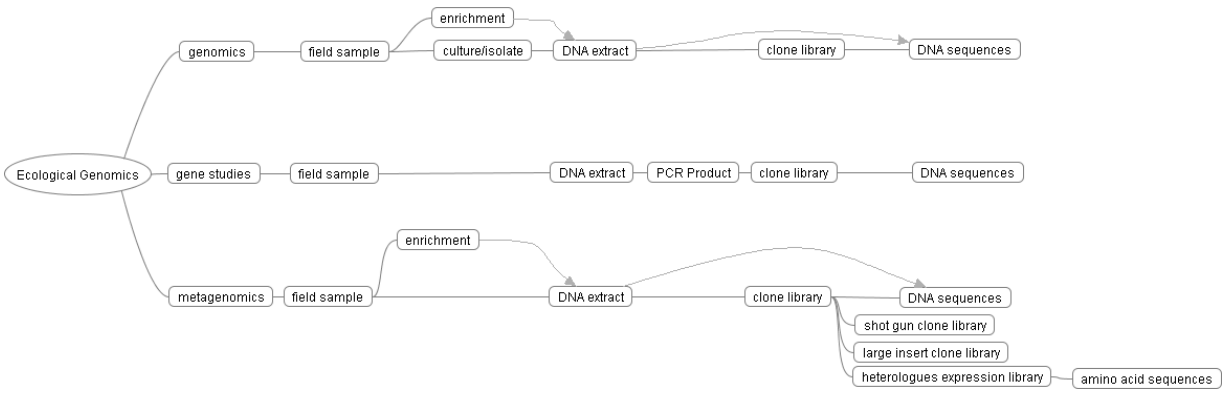

Figure 1 Samples produced within the workflow of typical ecological genomics study approaches.

price and the name of the study or the funding agency are, from this perspective, irrelevant although they would possibly fit into this model.

The example of a field sample is also taken to again emphasize that this is the central entity to which data available in environmental databases logically converges, provided that it is properly geo-referenced. This outlined domain model seems to provide the unifying schema for the integration of molecular sequence and environmental data because it reflects the natural hierarchy that exists among samples and information and materials derived from them, and maintains the complete research context.

Research Aims

The primary use of a (integration[10]) system is to gain some knowledge from large amount of data to then formulate hypothesis from the knowledge acquired, and finally perhaps to validate these hypothesis. – Hernandez 2007

The advances in ecological genomics demand an integrative framework for data analysis and knowledge generation (Straalen and Roelofs 2006). Primarily driven by advancements in sequencing technologies and the growing interest in metagenomics, an unprecedented mass of sequence data has become available and is continuously growing. Based on this development it is now possible to gain insight into whole microbial community structures and functions on the genomic level of any environment. Coupling these data with environmental data allows performing ecologically motivated comparative studies of genomes and metagenomes. Nevertheless, systematic, reproducible, and standardized comparative genomic studies are only possible if the molecular sequence- and environmental data are integrated into a single system.

The overall aim of this thesis is to develop a consistent data-analysis framework that accumulates algorithms and tools supporting ecological research questions. Such a system should allow the exploration and analysis of molecular sequence data in an environmental context. It should, furthermore, be made publically available and allow users to systematically explore the sequence space from a geographic perspective. The system should also deliver data relevant to questions like “Who is out there and where?” in terms of sequenced genomes and key genes, “What are they doing?” in terms of functional capacities, “under which environmental conditions?” and “what is the community structure?” in terms of gene fingerprints.

The first objective of this thesis is to develop and implement a database as well as a software architectural concept to integrate geographic, environmental and molecular sequence data. This system should have the following features:

1. A clear domain model
2. A clear rational and selection of relevant data
3. To be based on data and software standards

[10] Added by author for clarity

4. To be capable to store terabytes of data
5. Integrate data from a wide variety of sources
6. Software components for the exploration and analysis of the integrated data

The second objective is to find new ways of acquisition and integration of such data.

Thirdly, the exploitation of the integrated data by scientists should be facilitated by

a) supporting direct access to the integrated data, and
b) addition of data as demanded by the research questions at hand.

This aims to, in first place, gain new insights by the validation of hypothesis which cannot be answered with existing software tools. Additionally, such a system would demonstrate the great value of the integrated data to the ecological genomic research community.

Results and Discussion

Overview

Overview of scholarly published results:

I

Megx.net – database resources for marine ecological genomics.

Authors: Thierry Lombardot, Renzo Kottmann, Hauke Pfeffer, Michael Richter, Hanno Teeling, Christian Quast, and Frank Oliver Glöckner.

Published in Nucleic Acids Research. 2006; 34(suppl_1): D390-393.

Contribution: Architectural layout, MegDb database, integrated data, web pages.

II

Megx.net: integrated database resource for microbial ecological genomics

Authors: Renzo Kottmann, Ivalyo Kostadinov, Melissa B. Duhaime, Gregory Giuliani, Andrea de Bono, Anthony Lehmann, Frank Oliver Glöckner

Manuscript

III

MetaLook: a 3D visualisation software for marine ecological genomics.

Authors: Thierry Lombardot, Renzo Kottmann, Gregory Giuliani, Andrea de Bono, Nans Addor, and Frank Oliver Glöckner.

Published in *BMC Bioinformatics*. 2008; 8(1): 406.

Contribution: Use case definition, design, database.

IV

MetaMine - A tool to detect and analyse gene patterns in their environmental context

Authors: Uta Bohnebeck, Thierry Lombardot, Renzo Kottmann, and Frank Oliver Glöckner

Published in *BMC Bioinformatics*. 2008; 9(1): 459.

Contribution: Use case definition, algorithm outline, system architecture, database.

V

The minimum information about a genome sequence (MIGS) specification.

Authors: Dawn Field, George Garrity, Tanya Gray, Norman Morrison, Jeremy Selengut, Peter Sterk, Tatiana Tatusova, et al.

Published in *Nature Biotechnology.* 2008; 26(5): 541-547.

Contribution: MIMS extension.

VI

A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML)

Authors: Renzo Kottmann, Tanya Gray, Sean Murphy, Leonid Kagan, Saul Kravitz, Thierry Lombardot, Dawn Field, and Frank Oliver Glöckner. 2008.

Published in *Omics*. 2008; 12(2): 101-8.

VII

Defining Linkages between the GSC and NSF's LTER Program: How the Ecological Metadata Language (EML) Relates to GCDML and Other Outcomes

Authors: Inigo San Gil, Wade Sheldon, Tom Schmidt, Mark Servilla, Raul Aguilar, Corinna Gries, Tanya Gray, Dawn Field, James Cole, Jerry Yun Pan, Giri Palanisamy, Donald Henshaw, Margaret O'Brien, Linda Kinkel, Katherine McMahon, Renzo Kottmann et al.

Published in *Omics.* 2008; 12(2):151-6.

Contribution: Evaluation and discussion of the relation between EML and GCDML.

VIII

Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata

Authors: Lynette Hirschman, Cheryl Clark, K. Bretonnel Cohen, Scott Mardis, Joanne Luciano, Renzo Kottmann, James Cole, et al.

Published in *Omics*. 2008; 12(2): 129-36.

Contribution: Term definition, result check, use case definition.

IX

A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes

Authors: Jörg Peplies, Renzo Kottmann, Wolfgang Ludwig, and Frank Oliver Glöckner

Published in *Systematic and Applied Microbiology*. 2008;31(4): 251-257.

Contribution: Metadata definition.

I: Megx.net – database resources for marine ecological genomics.

Authors: Thierry Lombardot, Renzo Kottmann, Hauke Pfeffer, Michael Richter, Hanno Teeling, Christian Quast, and Frank Oliver Glöckner.

Published in Nucleic Acids Research. 2006; 34(suppl_1): D390-393.

Contribution: Architectural layout, MegDb database, integrated data, web pages.

Nucleic Acids Research

Megx net—database resources for marine ecological genomics

Thierry Lombardot, Renzo Kottmann, Hauke Pfeffer, Michael Richter, Hanno Teeling, Christian Quast and Frank Oliver Glöckner

Nucleic Acids Res. 34:390-393, 2006.
doi:10.1093/nar/gkj070

D390–D393 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj070

Megx.net—database resources for marine ecological genomics

Thierry Lombardot[1], Renzo Kottmann[1], Hauke Pfeffer[1], Michael Richter[1], Hanno Teeling[1], Christian Quast[1] and Frank Oliver Glöckner[1,2,*]

[1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany and [2]International University Bremen, D-28759 Bremen, Germany

Received August 8, 2005; Revised and Accepted October 8, 2005

ABSTRACT

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

INTRODUCTION

Over the last decade microbiology has undergone several changes. Robert Koch's invention of pure culture techniques at the end of the 19th century focussed microbiology on the isolation of bacteria for laboratory studies. In 1987 Carl Woese introduced the ribosomal RNA as a stable molecular marker for the classification and identification of microorganisms (1). The 'winds of change' blew in the field of microbiology (2) when the first cultivation-independent investigations reported an immense array of completely unexpected microbial diversity in the environment (3). The landmark publication of the first complete genome sequence of *Haemophilus influenzae* in 1995 (4) has transformed biology into a massively parallel and high throughput endeavour. This 'genomic revolution' finally reached the field of marine ecological genomics in the year 2000, defined as: 'The application of genomic sciences to understanding the structure and function of marine ecosystems' (5). Since 1995, >260 microbial genomes have been fully sequenced, and 600 more are well on their way (5). While most projects focus on microorganisms of medical or biotechnological interest, 22 complete marine genomes of environmental organisms are already available, and ~130 marine isolates are currently sequenced (Moore foundation http://www.moore.org). Recently, this cultivation-based approach has been complemented by a number of groundbreaking cultivation-independent metagenomic studies, the most prominent being the Venter Sargasso Sea expedition in 2004 (6), delivering >1.2 million new genes. This wealth of information caused a quantum leap in marine sciences and demands for different kinds of databases to transfer information into knowledge (7). The sequences, genomes, genes and predicted metabolic functions can not longer be regarded in an organism centric view but have to be handled in the context of the environment surrounding them. Therefore, it is necessary to link any environmental sequence information with its geographical location. This allows to correlate

*To whom correspondence should be addressed. Tel: +49 0421 2028938; Fax: +49 0421 2028580; Email: fog@mpi bremen.de

Nucleic Acids Research, 2006, Vol. 34, Database issue **D391**

the genomic features found at a distinct sampling site with physical, chemical and biotic information to identify organism-specific adaptations and their role and impact on the environment. This new kind of integrative data resource opens the path to address questions like: Are there differences in the genetic repertoire when travelling from coastal marine sites to the open ocean? or Do habitat specific gene patterns with yet unknown functions exist? If the latter is true the correlation with site specific environmental parameters might allow predicting a potential function for them. Can these genetic properties in turn explain the distribution of the organisms?

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

SOURCES OF GENOMIC AND METAGENOMIC DATA

The genome sequences of all currently available marine microorganisms have been retrieved from the EMBL and GenBank databases (8,9). Twenty-two bacteria and archaea originating from the water column of the ocean and from marine sediments have been completely sequenced (October 2005). The sequences and associated gene annotation have been imported into a local relational database allowing fast data retrieval. The corresponding annotations originate from independent submissions to the EMBL or the GenBank databases, and are of variable quality owing to the following reasons: (i) the original annotations were performed at different times; (ii) no controlled vocabulary is used for gene product names; and (iii) the effort expended in assigning functions to genes is variable between genome projects. Ecologically relevant annotation highlights were selected from original genome publications for each organism.

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

GENOME BROWSING

The genome browser allows easy and fast access to the sequences, their geographical location and the annotation highlights of each marine microorganism in the database. For example, the unexpected archaea-like C1 metabolism genes found in the genome of *Rhodopirellula baltica* can be accessed in their genomic context by a simple mouse click (Figure 1). Fast text search in the original annotations and BLAST searches are also available.

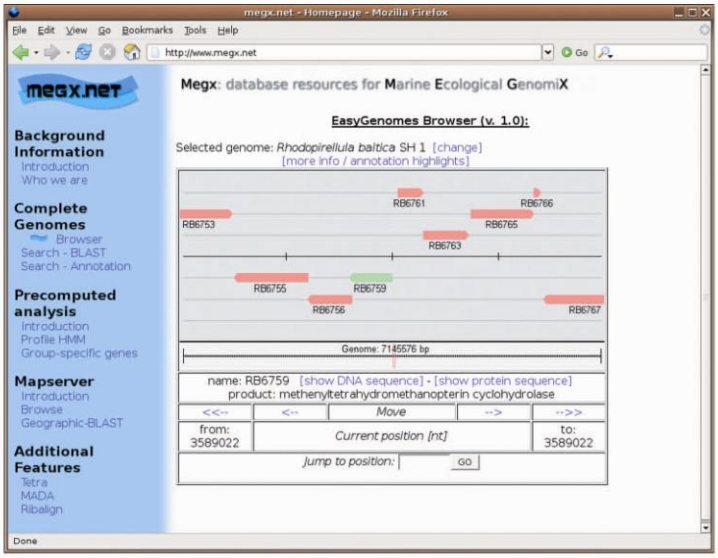

Figure 1. Fast access to the annotation highlights of marine microorganisms. Here, the archaea like C1 metabolism key gene is *R.baltica*.

PRECOMPUTED INFORMATION

Environmentally relevant protein families

Some gene families are of particular interest for ecological genomics, as they play key roles in the environment or give insights into the adaptation of microorganisms to their respective niche. Glycosylhydrolases, sulphatases, peptidases and transcriptional regulators are some examples of gene groups that have been automatically extracted based on selected profile hidden Markov models originating from the Pfam database (10). The results can be browsed graphically on our web page. This search strategy allows consistent quantitative comparisons, as the publicly available original annotation can not easily be compared. For example, the outstanding number of genes encoding sulphatases in the genome of *R.baltica* (11) or the reduced dataset of transcriptional regulators in *Prochlorococcus marinus* strains (12,13) can be compared with the corresponding gene content of other marine microorganisms.

Group-specific genes

Group-specific genes are defined as those found exclusively in a defined subset of genomes. The definition of groups is variable and can be based on a phylogenetic affiliation, a common metabolism or related habitats. An example for group specific genes for phylogenetically closely related organisms are the three available *P.marinus* strains. The results show that some light-inducible proteins are exclusively found in those organisms (13). Moreover, we present a set of proteins of yet unknown function which are *P.marinus* specific. The corresponding genes represent interesting targets for functional genomics and further wet-lab experiments.

TETRA SERVER

TETRA is a software tool for genomic and metagenomic analysis. It can assess the relatedness of genomic fragments by computing correlations between their tetranucleotide usage patterns (i.e. statistical over- and under-representation of tetranucleotides) (14,15). The new version includes chaos game plot representations for DNA sequences, which can be used to get additional information on the relatedness of genomic fragments. Moreover, TETRA can plot fluctuations of tetranucleotide usage patterns within DNA sequences. This is particularly useful to identify irregular regions in entire genomes or larger genomic fragments like laterally transferred genes or transposase and phage insertions.

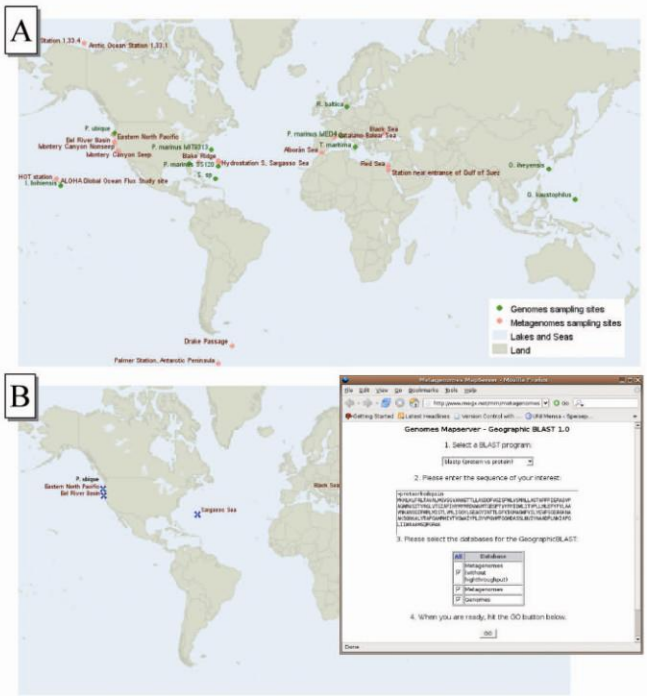

Figure 2. The Genomes Mapserver. (**A**) Marine genomes and metagenomic fragments can be browsed and searched on a world map on our web based system. (**B**) An example showing a Geographic BLAST search for genes encoding proteorhodopsins in the currently available dataset.

GENOMES MAPSERVER

Geographic information systems (GIS) are commonly used in the field of geology for data integration. A GIS is a combination of elements designed to store, retrieve, analyse and display geographic data. We introduce here the Genomes Mapserver, a GIS that allows access to genomic and metagenomic sequence data in their geographic and ecological contexts. The sampling sites of marine (meta)-genomic studies are displayed within a browsable world map (Figure 2). Each sampling site can be selected to display the corresponding sequences and additional contextual information. The underlying database is designed to enable future data mining tasks to reveal possible gene patterns associated with a particular environmental context. For targeted searches, a geographic-BLAST tool has been developed, allowing to perform 'spatial' queries for sequences based on the popular BLAST algorithm (16). The Geographic-BLAST/Genomes Mapserver combination allows to systematically study the biogeography of particular genes in the environment (Figure 2).

ADDITIONAL FEATURES

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

DATABASES ACCESS

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

ACKNOWLEDGEMENTS

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

Conflict of interest statement. None declared.

REFERENCES

1. Woese,C.R. (1987) Bacterial evolution. *Microbiol. Rev.*, **51**, 221 271.
2. Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.*, **176**, 1 6.
3. Torsvik,V., Goksoyr,J. and Daae,F.L. (1990) High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.*, **56**, 782 787.
4. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole genome random sequencing and assembly of *Haemophilus influenzae*. *Science*, **269**, 496 512.
5. Cary,C. and Chisholm,P. (2000) *Report of a Workshop on Marine Microbial Genomics to Develop Recommendations for the National Science Foundation*. Arlington, VA.
6. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D.Y., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66 74.
7. DeLong,E.F. and Karl,D.M. (2005) Genomic perspectives in microbial oceaonography. *Nature Insight*, **437**, nature04157.
8. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acid Res.*, **33**, D29 D33.
9. Benson,D.A., Karsch Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acid Res.*, **33**, D34 D38.
10. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acid Res.*, **32**, D138 D141.
11. Glöckner,F.O., Kube,M., Bauer,M., Teeling,H., Lombardot,T., Ludwig,W., Gade,D., Beck,A., Borzym,K., Heitmann,K. *et al.* (2003) Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc. Natl Acad. Sci. USA*, **100**, 8298 8303.
12. Dufresne,A., Salanoubat,M., Partensky,F., Artiguenave,F., Axmann,I.M., Barbe,V., Duprat,S., Galperin,M.Y., Koonin,E.V., Le Gall,F. *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl Acad. Sci. USA*, **100**, 10020 10025.
13. Rocap,G., Larimer,F.W., Lamerdin,J., Malfatti,S., Chain,P., Ahlgren,N.A., Arellano,A., Coleman,M., Hauser,L., Hess,W.R. *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, **424**, 1042 1047.
14. Teeling,H., Meyerdierks,A., Bauer,M., Amann,R. and Glöckner,F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, **6**, 938 947.
15. Teeling,H., Waldmann,J., Lombardot,T., Bauer,M. and Glöckner,F.O. (2004) TETRA: a web service and a stand alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403 410.

II: megx.net: integrated database resource for microbial ecological genomics

Authors: Renzo Kottmann, Ivalyo Kostadinov, Melissa B. Duhaime, Gregory Giuliani, Andrea de Bono, Anthony Lehmann, Frank Oliver Glöckner

Manuscript

Megx.net: integrated database resource for microbial ecological genomics

Renzo Kottmann[a,b], Ivalyo Kostadinov[a,b], Melissa B. Duhaime[a,b], Gregory Giuliani[c], Andrea de Bono[c], Anthony Lehmann[c], Frank Oliver Glöckner[a,b,*]

[a]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

[b]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

[c]GRID-Europe International Environment House, 1219 Châtelaine, Switzerland

[*]to whom correspondence should be addressed

Frank Oliver Glöckner

Max Planck Institute for Marine Microbiology

Celsiusstrasse 1

D-28359 Bremen, Germany

Phone: +49 421 2028970

FAX: +49 421 2028580

Email: fog@mpi-bremen.de

Keywords: genomics, metagenomics, environmental data, database, mapserver

Abstract

The megx.net database is a comprehensive resource that provides integrated, georeferenced information on genome and metagenome projects for microbial ecological genomics. All data are stored in the Microbial Ecological Genomics DataBase (MegDB), with its subdivisions MetaStorage for ‘on site’ sequence and habitat data, and OceaniaDB for global environmental data layers. The extended system provides access to several hundreds of genomes and metagenomes from prokaryotes and viruses. With the refined Genes Mapserver, all data can be interactively visualized on a world map and ‘on the fly’ statistics describing the fluctuation of environmental parameters for every sampling site can be calculated. Sequence entries have been curated to comply with the proposed minimal standards for genomes and metagenomes of the Genomic Standards Consortium (MIGS). Programmatic exchange of data is facilitated by Web Services. The integration of additional molecular diversity data at each sampling site is underway. A set of new tools for data analysis and visualization is available from the webpage, where all resources are freely accessible: http://www.megx.net.

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 the environmental, specifically the marine, sector that currently delivers the highest quantity of data. Marine ecosystems, covering more than 70% of the earth's surface, host the majority of biomass and significantly contribute to global organic matter and energy cycling. Microorganisms are known to be the "gatekeepers" of these processes and any insights into their life style 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 campaign (1,2) and Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project (http://www.moore.org/microgenome/). Notably, the Global Ocean Sampling resulted in a major single input of new sequence data with unprecedented functional diversity (3). The resulting flood of sequence data available in public databases is an extraordinary resource to explore the 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 calling, and annotation – fundamental steps in genomic analysis. Several new dedicated database resources have recently emerged to tackle the current needs for large scale metagenomic data management, namely, CAMERA (4), IMG/M (5), and RAST (6).

Nevertheless, it is becoming increasingly apparent that the full potential of comparative genome and metagenome analysis can be achieved only if the sequence data is considered in light of its original geographic and environmental context (7,8). The metadata describing a sample's geographic location and habitat, the details of its processing, from the time of sampling up to sequencing, and subsequent analyses, are important for e.g. modeling species' responses to environmental changes, or the spread and niche adaptation of bacteria and viruses. This suite of metadata is collectively referred as contextual data (9).

The megx.net portal for microbial ecological genomics was the first database to provide access to geographically integrated information on microbial genes and genomes in their marine environmental context (10). In addition to storing all 'on site' data describing sampling time, location, and field measurements of genomic sampling events, the extended

megx.net database resource now allows post factum retrieval of interpolated environmental parameters, such as temperature, pH, 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 all genomes and draft genomes available for prokaryotes and marine viruses. Megx.net is fully compliant with the Minimum Information about a Genome Sequence (MIGS) standard and its extension, Minimum Information about a Metagenome Sequence (MIMS) (7,9). Furthermore, an extended set of tools provides targeted access to the database content.

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 (1) MetaStorage, which stores georeferenced DNA sequence data from a collection of sequences from genomes, metagenomes, and genes of molecular environmental surveys, and associated contextual data, and (2) OceaniaDB, which stores georeferenced quantitative environmental data. The MetaStorage sequences are retrieved from the databases of the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/), are further supplemented by contextual data from GOLD (11) and NCBI Genome Projects (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html), and are then manually curated. Currently, it hosts draft and complete genomes, marine virus genomes, marine shotgun metagenomic datasets such as GOS, and large insert metagenomic datasets. An overview about the components and databases in mgex.net can be found in Figure 1.

Extension to draft genomes and shotgun datasets

The advances in sequencing technology have resulted in, an increasing number of genome and metagenome sequencing projects that are currently in progress, or stalled in a draft status

(11). In January 2008, GOLD (11) reported 3520 genome projects, of which less than one thousand were finished. Thus, most of the sequenced functional diversity is contained in these draft and shotgun datasets. To accommodate for this situation, megx.net was extended to host draft genomes and whole genome shotgun (WGS) datasets as well.

Extension to marine viruses

At an estimated ~10^{30} viruses in the oceans (12), viruses are increasingly recognized as the most abundant biological entity on the planet, the majority of which are bacteriophages, viruses that infect bacteria. Some of the first marine phage genomes to be sequenced revealed photosynthesis genes, not only rampantly transferred between phage and host, but also expressed during host infection (13-15). Furthermore, analysis of the 'viral' classified scaffolds from the Global Ocean Survey dataset has identified a plethora of host-specific, environmentally significant functional genes, including genes involved in photosynthesis, phosphate stress response, vitamin biosynthesis, antibiotic resistance, and nitrogen fixation (16). Considering their abundance and potential metabolic impact on the world's oceans, and a community call for integration of genomic and biogeochemical data (17), 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 bacteriophage genomes in MegDB. When sufficient sample coordinate, depth, and time data was found, the genomes are accompanied by interpolated contextual data available through the Genes Mapserver that was missing from the original genome publications.

Extension to MIGS compliant data

All datasets are georeferenced by the addition of longitude, latitude, depth/altitude and time (x, y z, t) of the sequenced samples. When possible, additional contextual data, such as habitat parameters measured in the field, are manually added and curated. This new release includes a habitat classification for 1850 available genomes using the Habitat-Lite ontology (18).

MetaStorage is designed to store all contextual data recommended by the Genomics Standards Consortium (GSC), and thus compliant with the MIGS standard, and its extension, MIMS (7). In addition, megx.net provides read-only access to MIGS/MIMS reports in Genomic Contextual Data Markup Language (GCDML) XML files. GCDML is a core

project of the GSC and is an XML Schema for generating MIGS/MIMS compliant reports for data entry, exchange, and storage (9).

New OceaniaDB: Additional environmental data

To supplement georeferenced molecular and 'on site' data with interpolated, environmental parameters, OceaniaDB was added to MegDB. OceaniaDB is an 'oceanographic/environmental' database consisting of essential environmental data layers for aquatic ecosystems. It provides physical, chemical, geological and biological parameters, such as ocean water temperature and salinity, nutrient concentrations, organic matter and chlorophyll.

The different layers comprise information from three sources:

- World Ocean Atlas: a set of objectively analyzed (one decimal degree spatial resolution) climatological fields of *in situ* measurements (http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html);
- World Ocean Database: a collection of scientific, quality-controlled ocean profiles (http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html);
- SeaWIFS chlorophyll *a* data (http://seawifs.gsfc.nasa.gov).

These layers are described at 33 standard depths for annual, seasonal and monthly periods, such that the geographic data (x, y z, t) functions as a universal anchor between layers. Facilitated by the PostGIS extension of MegDB, the environmental layers of OceaniaDB are integrated with the contextual data recorded at the time of sampling and stored in MetaStorage. 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.

Genes Mapserver

The Genes Mapserver (formerly Metagenomes Mapserver) offers a sample-centric view of the MetaStorage content, showing the sampling sites of genomes, metagenomes and single genes from sequencing surveys. Substantial improvements to the underlying Geographic Information System (GIS) and web view were made. The website is now interactive, offering Google Maps-like navigation and an overlay of sampling sites on the world maps

representing the OceaniaDB environmental data layers. Sample site details can be retrieved by clicking the sampling points on the map.

The modular web application and database are based on Open Source software. The Genes Mapserver allows extraction of interpolated values for several physico-chemical and biological parameters, such as temperature, dissolved oxygen, nitrate and chlorophyll concentrations, over specified time intervals (month-long, annual, or seasonal).

The addition of georeferenced small subunit ribosomal RNA (rRNA) sequences from the SILVA rRNA databases project (19) to the Genes Mapserver introduces a first approach towards the integration of microbial diversity with specific sampling sites. Although only roughly 5% of the nearly 700,000 sequences in SILVA SSUParc database are georeferenced at the moment, efforts are ongoing (http://www.arb-silva.de/projects/contextual-data/) to significantly increase this situation for the future.

Geographic-BLAST

The Geographic-BLAST tool queries the MegDB genome and metagenome sequence data using the BLAST algorithm (20). The results are reported according to the original sample locations of the database hits. With the updated Geographic-BLAST, results are plotted on four world map views organized by (1) samples, differentiating metagenome from genome hits, (2) e-value intervals, (3) number of hits per sample site, and (4) the percentage of all coding sequences (CDS) at a sample site with hits. A result output table also displays useful information about the BLAST hits. An additional table displays useful information about the BLAST hits.

Extended Toolbox for Data Analysis

Recently, new tools were developed that offer diverse methods to access and analyze the MegDB content, and are now available at megx.net.

MetaLook offers a gene-centric view of the MetaStorage content, with a special focus on habitat parameters ((21), http://www.megx.net/metalook). It is a desktop application with a 3D user interface to interactively visualize DNA sequences on a world map. The user can define environmental containers to organize sequences according to different habitat criteria.

These sequence sets can be queried by Geographic-BLAST with either genes in the database or user-imported sequences. This allows an interactive assessment of the distribution of gene functions in the environment.

MetaMine is an interactive data mining tool which enables the detection of gene patterns in an environmental context ((22), http://www.megx.net/metamine). This desktop application offers a targeted, semi-automatic search for gene patterns based on user expertise. MetaMine implements a client/server architecture to both perform BLAST searches against, and retrieve environmental data from, MegDB. The user-friendly graphical user interface allows further inspection of calculated gene patterns in an ecological context.

JCoast is a desktop application primarily designed to analyze and compare (meta)genome sequences of prokaryotes ((23), http://www.megx.net/jcoast). JCoast offers a flexible graphical user interface (GUI), as well as an application programming interface (API) that facilitates back-end data access to GenDB projects (24). JCoast offers individual, cross genome and metagenome analysis, including access to Geographic-BLAST.

Web Services

The newly extended version of megx.net now offers Web Services to programmatically access the MegDB content. All geographical maps can be retrieved via simple HTTP GET requests as specified by the Web Map Service (WMS) standard. The base URL for WMS requests is http://www.megx.net/wms. The first sets of MIGS/MIMS compliant reports are available in XML format for direct download. Theses reports contain extensively manually curated contextual data encoded in the GCDML

Outlook

Although many new datasets have been included in the extended version of megx.net, many datasets in the public repositories still lack the minimal contextual data necessary for georeferenced data integration into MegDB. To improve this situation, megx.net promotes the adoption of existing and newly developed standards. It is likely that the genome catalogue, provided by the Genomic Standards Consortium, will serve as a prime source for georeferenced genomes and metagenomes in the future. In cooperation with the International Census of Marine Microbes (ICoMM), megx.net seeks to significantly enhance the integration of microbial diversity data. Furthermore, multivariate analysis tools to determine

the key habitat parameters triggering the ‘on site’ functional diversity are planned.

The megx.net database and tools are meant as an integrative resource for the analysis of microbial diversity and function on the molecular level with respect to habitat traits. A holistic view of the complex interplay of organisms, genes, and the environment surrounding them is a major step towards a better understanding of the complex responses and adaptations of organisms to environmental changes. The megx.net integrated datasets are a cornerstone for the emerging field of eco-system biology.

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.

Funding

This work was supported by the FP6 EU project MetaFunctions (grant CT 511784), the Network of Excellence “Marine Genomics Europe” and the Max Planck Society.

Acknowledgments

We would like to acknowledge Wolfgang Hankeln, Magdalena Golden, Pier Luigi Buttigieg, Mitul Jain, and Laura Sandrine for their valuable input to megx.net as well as David E. Todd for redesigning the web page.

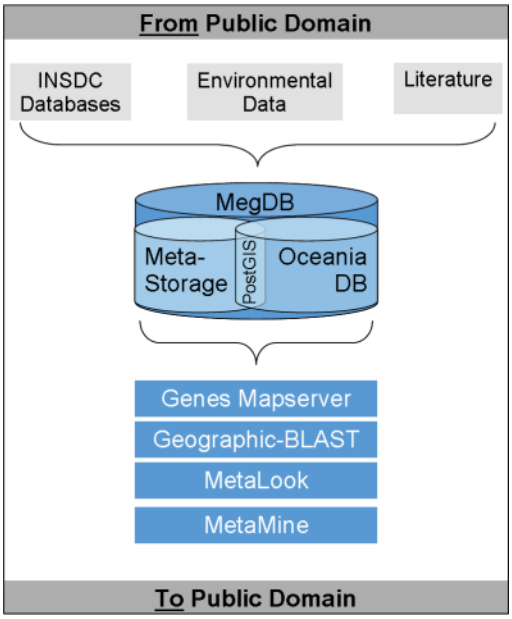

Figure Legend

Fig. 1: General Architecture of megx.net: DNA sequence Data is integrated with contextual data retrieved from diverse resources including manual and semi-automatic literature analysis. MegDB integrates the data conforming OGC standards and the MIGS/MIMS specification. Several external tools are available to access the MegDB content.

References:

1. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D.Y., Paulsen, I., Nelson, K.E., Nelson, W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66-74.
2. 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. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS. Biol.*, **5**, e77.
3. 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. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS. Biol.*, **5**, e16.
4. Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P. and Frazier, M. (2007) CAMERA: A Community Resource for Metagenomics. *PLoS. Biol.*, **5**, e75.
5. Markowitz, V.M., Ivanova, N.N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.M.A., Grechkin, Y., Dubchak, I., Anderson, I. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acid Res.*, **36**, D534-D538.

6. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M. *et al.* (2008) The RAST server: Rapid annotations using subsystems technology. *Bmc Genomics*, **9**.
7. Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M.J., Angiuoli, S.V. *et al.* (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.*, **26**, 541-547.
8. Field, D., Morrison, N., Glöckner, F.O., Kottmann, R., Cochrane, G., Vaughan, R., Garrity, G., Cole, J., Hirschman, L., Schriml, L. *et al.* (2008) Working together to put molecules on the map. *Nature*, **453**, 978-978.
9. Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., Glöckner, F.O. and Consortium, t.G.S. (2008) A standard MIGS/MIMS compliant XML schema: Toward the development of the Genomic Contextual Data Markup Language (GCDML). *Omics*, **12**, 115-121.
10. Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C. and Glöckner, F.O. (2006) Megx.net - database resource for marine ecological genomics. *Nucleic Acid Res.*, **34**, D390-D393.
11. 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 Acid Res.*, **36**, D475-D479.
12. Suttle, C.A. (2005) Viruses in the sea. *Nature*, **437**, 356-361.
13. Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M. and Chisholm, S.W. (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, **438**, 86-89.
14. 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**, 790-806.
15. Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F. and Chisholm, S.W. (2004) Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc. Natl. Acad. Sci. USA*, **101**, 11013-11018.
16. 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. *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, **3**, e1456.
17. 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. *et al.* (2008) Global-scale processes with a nanoscale drive: the role of marine viruses. *Isme J.*, **2**, 575-578.
18. Hirschman, L., Clark, C., Cohen, K.B., Mardis, S., Luciano, J., Kottmann, R., Cole, J., Markowitz, V., Kyrpides, N., Morrison, N. *et al.* (2008) Habitat-Lite: A GSC case study based on free text terms for environmental metadata. *Omics-a Journal of Integrative Biology*, **12**, 129-136.
19. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W.G., 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 Acid Res.*, **35**, 7188-7196.
20. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.
21. Lombardot, T., Kottmann, R., Giuliani, G., de Bono, A., Addor, N. and Glöckner, F.O. (2007) MetaLook: a 3D visualisation software for marine ecological genomics. *BMC Bioinformatics*, **8**.
22. Bohnebeck, U., Lombardot, T., Kottmann, R. and Glöckner, F.O. (in press) MetaMine - A tool to detect and analyse environmentally relevant gene patterns. *BMC Bioinformatics*.

23. Richter, M., Lombardot, T., Kostadinov, I., Kottmann, R., Duhaime, M.B., Peplies, J. and Glöckner, F.O. (2008) JCoast - A biologist-centric software tool for data mining and comparison of prokaryotic (meta) genomes. *BMC Bioinformatics*, **9**.
24. Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. *et al.* (2003) GenDB - an open source genome annotation system for prokaryote genomes. *Nucleic Acid Res.*, **31**, 2187-2195.

III. MetaLook: a 3D visualisation software for marine ecological genomics.

Authors: Thierry Lombardot, Renzo Kottmann, Gregory Giuliani, Andrea de Bono, Nans Addor, and Frank Oliver Glöckner.

Published in *BMC Bioinformatics*. 2008; 8(1): 406.

Contribution: Use case definition, design, database.

BMC Bioinformatics

BioMed Central

Software

Open Access

MetaLook: a 3D visualisation software for marine ecological genomics

Thierry Lombardot*[1], Renzo Kottmann[1,3], Gregory Giuliani[2], Andrea de Bono[2], Nans Addor[2] and Frank Oliver Glöckner[1,3]

Address: [1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany, [2]Division of Early Warning and Assessment, Global Resource Information Database – Europe, United Nations Environment Programme, International Environment House, 1219 Châtelaine, Switzerland and [3]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

Email: Thierry Lombardot* - tlombard@mpi-bremen.de; Renzo Kottmann - rkottman@mpi-bremen.de; Gregory Giuliani - giuliani@grid.unep.ch; Andrea de Bono - debono@grid.unep.ch; Nans Addor - nans.addor@gmail.com; Frank Oliver Glöckner - fog@mpi-bremen.de

* Corresponding author

Published: 22 October 2007

BMC Bioinformatics 2007, 8:406 doi:10.1186/1471-2105-8-406

Received: 30 May 2007
Accepted: 22 October 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/406

Abstract

Background: Marine ecological genomics can be defined as the application of genomic sciences to understand the structure and function of marine ecosystems. In this field of research, the analysis of genomes and metagenomes of environmental relevance must take into account the corresponding habitat (contextual) data, e.g. water depth, physical and chemical parameters. The creation of specialised software tools and databases is requisite to allow this new kind of integrated analysis.

Results: We implemented the MetaLook software for visualisation and analysis of marine ecological genomic and metagenomic data with respect to habitat parameters. MetaLook offers a three-dimensional user interface to interactively visualise DNA sequences on a world map, based on a centralised georeferenced database. The user can define *environmental containers* to organise the sequences according to different habitat criteria. To find similar sequences, the containers can be queried with either genes from the georeferenced database or user-imported sequences, using the BLAST algorithm. This allows an interactive assessment of the distribution of gene functions in the environment.

Conclusion: MetaLook allows scientists to investigate sequence data in their environmental context and to explore correlations between genes and habitat parameters. This software is a step towards the creation of specialised tools to study constrained distributions and habitat specificity of genes correlated with specific processes.

MetaLook is available at http://www.megx.net/metalook

Background

The cost reduction and high-throughput automation of DNA sequencing over the last years have had a profound impact on the field of microbial ecology, giving birth to the field of ecological genomics. Ecological genomics can be defined as the application of genomic sciences to

BMC Bioinformatics 2007, **8**:406 http://www.biomedcentral.com/1471-2105/8/406

understand the structure and function of marine ecosystems. This field of research is focussed on the investigation of environmentally relevant microorganisms taken from their natural habitats. The sequencing of the genomes of such organisms, especially the new wave of ecological metagenomics, in which DNA sequences are directly retrieved from the environment without prior cultivation, produces huge amounts of new proteins, which theoretically reflect the prominent metabolic processes in the environment [1,2].

Nevertheless, the functional potential coded in the DNA sequences can be successfully interpreted only if considered in their ecological context. Currently, general-purpose DNA databases, as provided by the International Nucleotide Sequence Database Collaboration (INSDC [3]), store only limited environmental contextual (meta-)information with the sequences, if any. Exact geographic origins and the corresponding on-site physical and chemical parameters are rarely found in these databases. This clearly hinders integrated ecological interpretations and limits the extraction of biological knowledge from raw sequence data. With the increasing awareness of this issue [1] and the introduction of new organisms and sample-centric contextual (meta)-data standards, such as those proposed by the Genomic Standards Consortium (GSC) [4,5], this is likely to change in the future. Furthermore, genomic and metagenomic sequence data can be supplemented by information extraction from the literature for proper georeferencing. In parallel, new specialised database architectures and software tools for data visualisation and interpretation are needed [6], enabling the representation of sequence and habitat data in a geographic information system [7,8]. Here we introduce MetaLook, a 3D visualisation software allowing browsing and interpretation of marine sequence data in their ecological context.

Implementation

Database server

Genomes and metagenomes from marine environments were selected for import from the NCBI databases [9] into a local PostgreSQL/PostGIS database [10], according to the following criteria: i) the DNA sequence must be of marine bacterial or archaeal origin; ii) sequence quality must be high (i.e. sequencing coverage of at least eight fold); iii) marker and single genes are rejected; and iv) the geographic origin of the DNA sequences must be known precisely (e.g. from the original publication). Lower quality sequences (draft genomes and short metagenomics reads) will be included in future releases.

Geographic locations were stored in our database for accepted DNA samples. Moreover, on-site contextual (meta-)data, such as physical and chemical parameters at the sampling site, were retrieved manually from the original publications and additional web pages when available. This manual curation step is crucial in order to reliably link on-site contextual data to DNA sequences. Moreover, having the exact geographic position for each sample in our database allows the interpolation of environmental parameters from worldwide data sets. Currently, the following global oceanic physical and chemical parameters are integrated into our database from the WOA data set (World Ocean Atlas): temperature, nitrate, phosphate, oxygen and silicate concentration, as well as salinity [11].

Java 3D-based client

The MetaLook interface is a locally running client based on the Java 3D API [12], started using the Java Web Start technology from the megx.net data portal [7]. The starting point of the interface is a 3D workbench displaying a world map with the sampling sites of genomic and metagenomic studies available in our database (Fig. 1). The 3D approach allows displaying larger amounts of data and interconnections than a classical 2D visualisation [13]. Within MetaLook, the user can sort the corresponding DNA sequence data into so-called *environmental containers*, which are flexible entities grouping data according to specific criteria, such as habitat types, ocean water depth or physical and chemical parameters. This custom data classification allows the user to define ecological niches according to specific biological questions (Fig. 2). The DNA sequence fragments grouped into the containers can be visualised on the workbench for browsing and comparing (Fig. 3). Moreover, each container can be searched for specific genes based on their annotations.

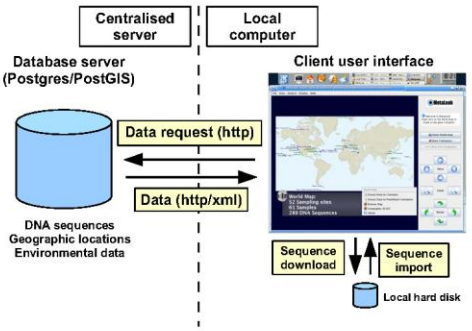

Figure 1
Client/Server architecture of MetaLook. The Java3D client runs on a local machine and gets data from the PostgreSQL server through HTTP request in XML format. DNA sequences of interest can be up- and downloaded for further analysis.

BMC Bioinformatics 2007, **8**:406 http://www.biomedcentral.com/1471-2105/8/406

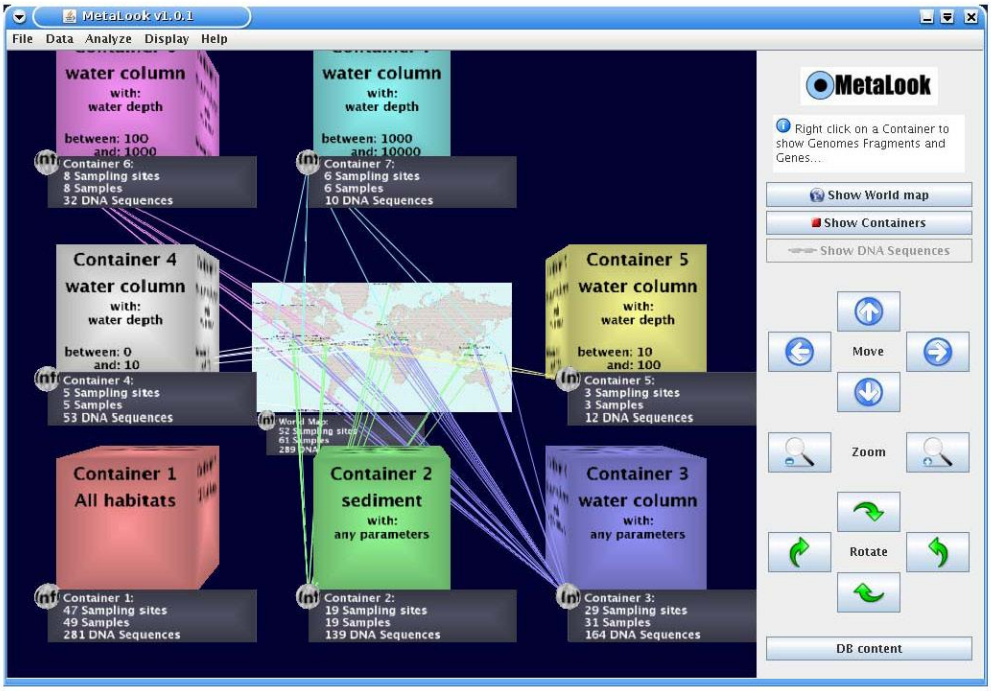

Figure 2
The *environmental containers* in MetaLook. DNA sequences of genomes and metagenomes can be sorted into 3D containers according to habitat information such as e.g. water column vs. sediments, depth profile or physical-chemical parameters. The geographic origins of the DNA sequence samples in each container are shown on the world map.

Search results are shown graphically in their genomic context. The DNA or protein sequence of each gene can be displayed or easily downloaded from the database. All DNA sequences in a container can be downloaded in batch mode. Custom sequences can be imported into the MetaLook interface in FASTA format.

BLAST against environmental containers

Any protein encoding gene from our georeferenced database or user-imported sequences can be used as a query for a BLASTP run [14] against the genes grouped into user-defined *environmental containers*. The BLASTP analysis is started from the MetaLook interface (client) and runs on the centralised server. The results are shown graphically using 3D connectors between the query gene and the containers with sequence matches (Fig. 4). This representation reveals the distribution of similar genes in the user-defined habitats. The results are saved in a result panel for detailed investigation, showing the habitat parameters of each match, the corresponding BLASTP *e*-value, and sequence alignment (Fig. 5a, b).

Comparison to other programs

Some interesting DNA sequence tools making use of 3D are currently available. Sockeye is a 3D environment for comparative genomics allowing simultaneous visualisation of the annotations of different eukaryotic organisms [15]. The Correlogo server is a tool to display DNA sequence alignments using 3D sequence logos [16]. The Walrus graph visualisation tool allows visualisation of very large phylogenetic trees in a hyperbolic space [17]. These examples show the benefits of advanced visualisation tools for DNA research and the management of large data sets. However, within this context, MetaLook is unique in its orientation toward environmental genomics, geographic and contextual data integration.

BMC Bioinformatics 2007, **8**:406 http://www.biomedcentral.com/1471-2105/8/406

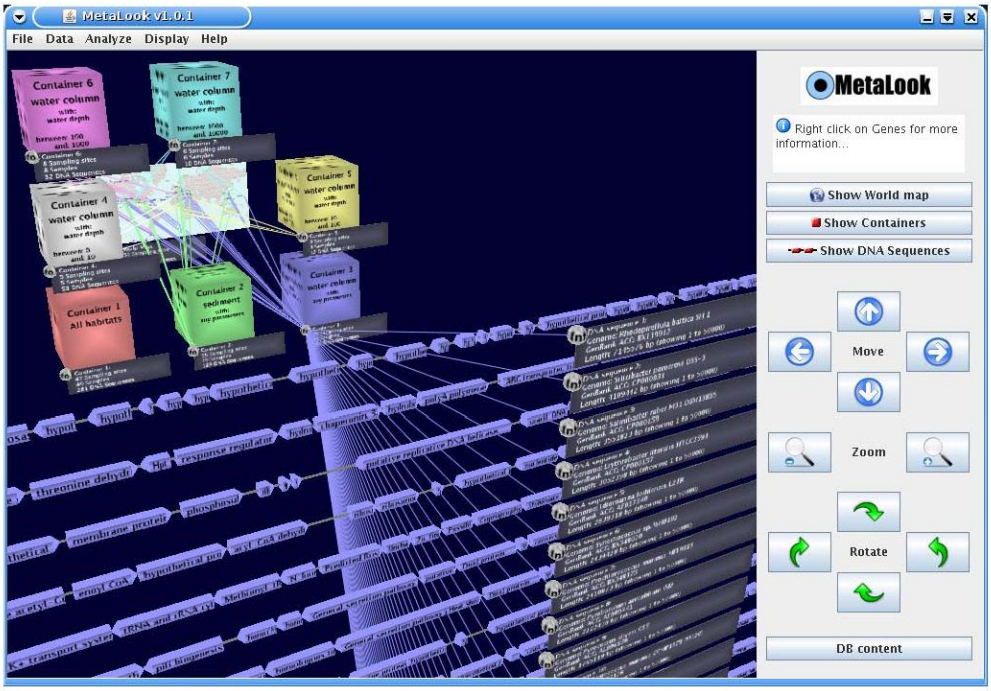

Figure 3
Displaying (meta)-genomes and genes in MetaLook. Each container can be opened to display the DNA sequence and the genes of each genome and metagenomic fragment. Genes can further be selected for download or analysis.

Results and Discussion

The MetaLook interface allows the sorting of sequence data according to sampling sites and habitat parameters, with respect to targeted biological questions. The distribution of genes in the environment is revealed using the BLAST algorithm with a selected query gene against other sequences sorted in *environmental containers*. The following examples illustrate some expected and unexpected habitat distributions of genes in the environment using the MetaLook interface.

Methanogenesis genes (mch and mcr)

In microorganisms, methanogenesis is a form of microbial anaerobic respiration leading to the formation of methane. Recent experimental and genomic data support the hypothesis that anaerobic oxidation of methane (AOM) is using a reverse-methanogenesis pathway [18-20]. Such biochemical processes are crucial in the environment, as methane is an important greenhouse gas contributing to global warming. One of the key genes of methanogenesis and AOM is mcr, encoding a methyl-coenzyme-M reductase (Mcr). The distribution of mcr in the environment was visualised by MetaLook with the following steps: i) predefined *environmental containers* were created from the world map, grouping sediment and water samples by depth (Fig. 2); ii) a text search for the gene "mcr" was performed; iii) Mcr protein sequences (e.g. McrB, [Genbank: AAB98847]) were blasted against all containers (BLASTP, *e*-value cut-off 10^{-10}). The results show that within the georeferenced marine bacteria and archaea currently available in our database, genes encoding Mcr are only found in sediments. Although expected, this observation shows that mcr genes are habitat specific, which is consistent with the strictly anaerobic nature of methanogenesis and the AOM process.

Another key gene of the methanogenesis and AOM processes is mch, encoding a methenyl-tetrahydromethanopterin cyclohydrolase (Mch). Interestingly, this gene was reported in some proteobacteria and planctomycetes,

BMC Bioinformatics 2007, **8**:406 http://www.biomedcentral.com/1471-2105/8/406

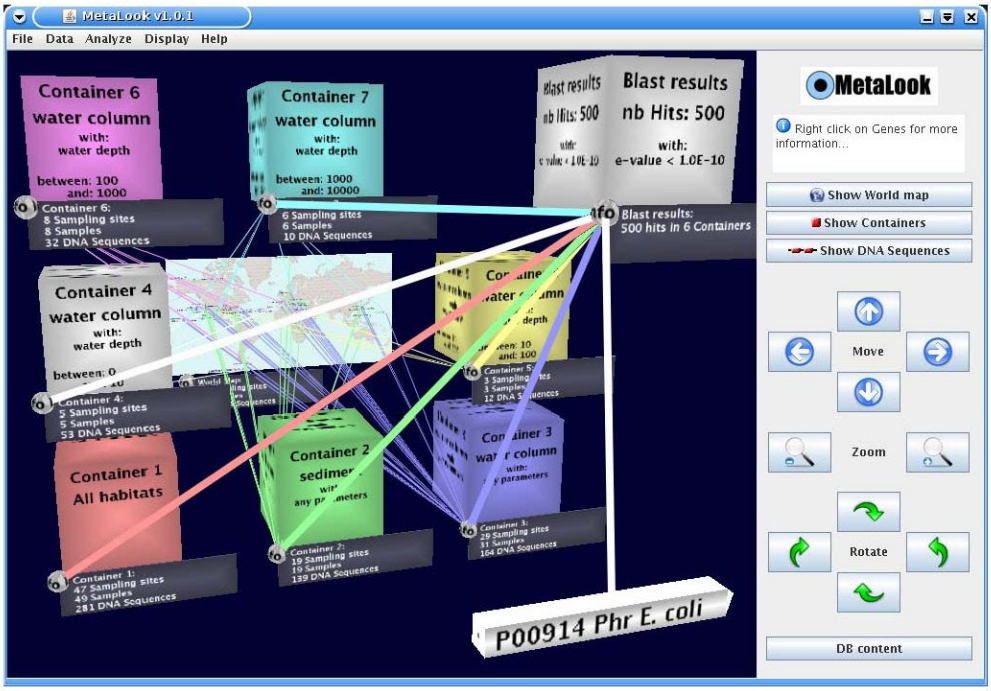

Figure 4
Study of the habitat-specificity of a gene. Here, the gene encoding a photolyase (foreground) shows BLASTP hits in the top layers of the ocean, as expected, but also some unexpected hits in the deep sea (container 7).

where an archaea-like C1 metabolism appears to be present [21]. Following the same procedure used for mcr (see above) revealed, as expected, that the mch gene is not only present in genomes and metagenomes originating from sediments, but is also found in the genome of at least one sea water column bacterium, the planctomycete *Rhodopirellula baltica* SH 1^T from the Baltic Sea [22] (e.g. [GenBank: CAD74990]). Furthermore, this analysis showed that mch is also found in the high-throughput metagenomics data set of the Sargasso Sea [23], suggesting an even more widespread distribution of this gene in the environment. Hence, the analysis of the habitat specificity of mcr and mch revealed differential environmental distribution of genes relevant for major biochemical processes involved in the global cycling of carbon.

Photolyase gene (phr)

Solar UV-light induces pyrimidine dimers in genomic material, leading to enhanced mutation rates. Photolyases are proteins involved in a light-dependant, single-step DNA repair mechanisms, which protect microorganisms against this destructive effect [24]. Comparative analysis of the genomes of three *Prochlorococcus marinus* strains, one of the most abundant phototrophic prokaryote in the ocean, previously reported the presence of photolyase encoding genes (phr) in the high-light ecotype, and its absence in the low-light ecotypes (water depth: 5 m and 120/135 m, respectively) [25]. This finding suggests that for this particular species, the phr gene is lost if an organism is exposed to little or no UV-light. As no DNA pyrimidine dimers should form where no UV-light stress occurs, the phr gene is not expected in the deep layers of the ocean.

To systematically test the occurrence of the phr gene in the marine environment, a phr gene with experimental evidence (*Escherichia coli* K-12, [Swiss-Prot: P00914]) was imported into the MetaLook interface and searched

BMC Bioinformatics 2007, **8**:406 http://www.biomedcentral.com/1471-2105/8/406

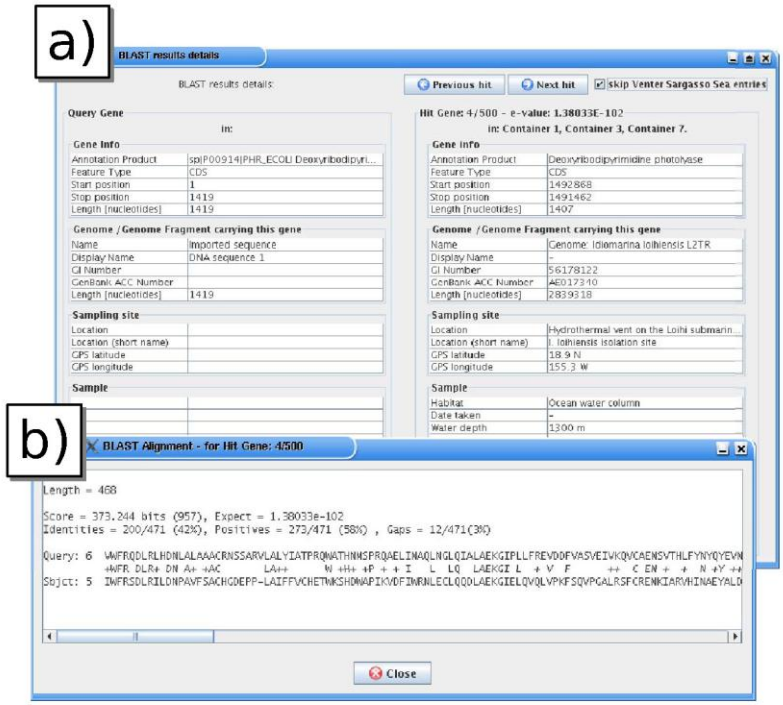

Figure 5
Study of the habitat-specificity of a gene (habitat parameters). a) Information for an unexpected BLASTP hit of the photolyase gene from figure 4 with a sequence originating from a deep-sea sample; b) BLASTP sequence alignment for the corresponding sequences.

against predefined environmental containers with the BLASTP algorithm (*e*-value cut-off 10^{-10}). Some sequence hits in the top layers of the ocean were found, as expected (e.g. *Prochlorococcus marinus* MED4,[Genbank:CAE18744] and *Rhodopirellula baltica* SH 1^T, [Genbank: CAD77347]). Moreover, unexpected sequences from deep-sea water (hot vent) and coastal sediments were also hit by this analysis (*Idiomarina loihiensis* L2TR, [GenBank: AAV82228] and *Hahella chejuensis* KCTC 2396, [GenBank: ABC28582]) [26,27] (Fig. 3). These genes are likely to be functional, with full-length BLASTP alignments and excellent statistical support, with *e*-values below 10^{-100} (Fig. 4a, b). Such unexpected occurrence of genes encoding photolyases in these environments might be explained by: i) the presence of allochthonous organisms [28], ii) residual phr genes awaiting deletion in organisms recently adapted to deep-sea or sediment environments, or iii) the possible need for protective mechanisms against geothermal light, even if the dominant wavelengths are not in the UV range [29].

Future work

The availability of worldwide physical and chemical parameters linked to DNA sequences opens the way to multivariate analysis. This approach will be crucial as more georeferenced genomic and metagenomic samples become available. The integration of low quality sequences (e.g. single reads from metagenomics) and biodiversity markers (e.g. ribosomal RNA genes) in our geographic-centric system is also a follow-up perspective.

Conclusion

Marine ecological genomics is an emerging field of research but available high quality and accurately georeferenced sequence data are still sparse compared to the natural habitat and organism diversity. Therefore, the observed absence of genes in particular habitats may reflect a mere gap in the database coverage. However, with the use of appropriate software tools, common knowledge can be easily confirmed and unexpected findings can be obtained for further investigation, as shown here with the example of a light-dependant gene present in the deep-sea. As more sequences with rich contextual (meta-) data from marine genome and metagenome projects are released, the accuracy and reliability of correlations between gene occurrence and habitat parameters will continuously improve. Targeted studies of gene distribution in the environment are greatly facilitated by our specialised databases and software tools presented here, offering an advanced software workbench for biologists.

Availability and requirements

Project name: MetaLook

Project home page: http://www.megx.net/metalook

Direct download and installation (Java web start): http://www.megx.net/metalook/MetaLook_start.jnlp

Operating systems: Windows or Linux.

Programming language: Java.

Other requirements: Java JRE 1.5 or higher, 3D card recommended.

License: license-free.

Any restrictions to use by non-academics: MetaLook may not be sold or bundled with any type of commercial application.

List of abbreviations used

mcr/Mcr: methyl-coenzyme-M reductase gene/protein.

mch/Mch: methenyl-tetrahydromethanopterin cyclohydrolase gene/protein.

phr/Phr: photolyase gene/protein.

FP6: the Sixth Framework Programme of the European Union.

NEST: new and emerging science and technology.

Competing interests

The author(s) declares that there are no competing interests.

Authors' contributions

TL designed and implemented MetaLook, the initial version of the underlying database and integrated the genomic data. RK designed and implemented the current version of the underlying database and integrated the metagenomic data. GG, AB and NA performed WOA data set integration and interpolations. FOG is leading the EU-project MetaFunctions, gave advise for software development, and has made revisions and contributions to the manuscript.

Acknowledgements

We thank the EU Sixth Framework Programme (FP6-NEST) for providing financial support (MetaFunctions project, contract no. 511784). We also thank Dr. Johanna Wesnigk for her management work within the MetaFunctions project and Melissa Duhaime for proofreading the manuscript. All authors read and approved the final manuscript. Funding to pay the Open Access publication charges for this article was provided by the Max Planck Society.

References

1. Field D, Kyrpides N: **The Positive Role of the Ecological Community in the Genomic Revolution.** *Microb Ecol* 2007, **53**:507-511.
2. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.** *PLoS Biol* 2007, **5**:e77.
3. **International Nucleotide Sequence Database Collaboration** [http://www.insdc.org]
4. **The Genomic Standards Consortium (GSC)** [http://darwin.nox.ac.uk/gsc/gcat]
5. Morrison N, Cochrane G, Faruque N, Tatusova T, Tateno Y, Hancock D, Field D: **Concept of sample in OMICS technology.** *OMICS* 2006, **10**:127-137.
6. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: A Community Resource for Metagenomics.** *PLoS Biol* 2007, **5**:e75.
7. Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glöckner FO: **Megx.net – database resources for marine ecological genomics.** *Nucleic Acids Res* 2006, **34**:D390-393.
8. **The Genomes Mapserver: a geographic information system for metagenomic and genomic sequences** [http://www.megx.net/gms]
9. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007, **35**:D5-12.
10. **PostGIS: support for geographic objects to the PostgreSQL object-relational database** [http://www.postgis.org]
11. **National Oceanographic Data Center (NODC) – World Ocean Atlas** [http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html]
12. **Java 3D API project homepage** [https://java3d.dev.java.net]

13. Bohannon J: **Bioinformatics. The human genome in 3D, at your fingertips.** *Science* 2002, **298**:737.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
15. Montgomery SB, Astakhova T, Bilenky M, Birney E, Fu T, Hassel M, Melsopp C, Rak M, Robertson AG, Sleumer M, Siddiqui AS, Jones SJ: **Sockeye: a 3D environment for comparative genomics.** *Genome Res* 2004, **14**:956-962.
16. Bindewald E, Schneider TD, Shapiro BA: **CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments.** *Nucleic Acids Res* 2006, **34**:W405-411.
17. Hughes T, Hyun Y, Liberles DA: **Visualising very large phylogenetic trees in three dimensional hyperbolic space.** *BMC Bioinformatics* 2004, **5**:48.
18. Kruger M, Meyerdierks A, Glöckner FO, Amann R, Widdel F, Kube M, Reinhardt R, Kahnt J, Bocher R, Thauer RK, Shima S: **A conspicuous nickel protein in microbial mats that oxidize methane anaerobically.** *Nature* 2003, **426**:878-881.
19. Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM, DeLong EF: **Reverse methanogenesis: testing the hypothesis with environmental genomics.** *Science* 2004, **305**:1457-1462.
20. Meyerdierks A, Kube M, Lombardot T, Knittel K, Bauer M, Glöckner FO, Reinhardt R, Amann R: **Insights into the genomes of archaea mediating the anaerobic oxidation of methane.** *Environ Microbiol* 2005, **7**:1937-1951.
21. Bauer M, Lombardot T, Teeling H, Ward NL, Amann RI, Glöckner FO: **Archaea-like genes for C1-transfer enzymes in Planctomycetes: phylogenetic implications of their unexpected presence in this phylum.** *J Mol Evol* 2004, **59**:571-586.
22. Glöckner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzym K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R: **Complete genome sequence of the marine planctomycete Pirellula sp. strain 1.** *Proc Natl Acad Sci USA* 2003, **100**:8298-8303.
23. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**:66-74.
24. Weber S: **Light-driven enzymatic catalysis of DNA repair: a review of recent biophysical studies on photolyase.** *Biochim Biophys Acta* 2005, **1707**:1-23.
25. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW: **Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation.** *Nature* 2003, **424**:1042-1047.
26. Hou S, Saw JH, Lee KS, Freitas TA, Belisle C, Kawarabayasi Y, Donachie SP, Pikina A, Galperin MY, Koonin EV, Makarova KS, Omelchenko MV, Sorokin A, Wolf YI, Li QX, Keum YS, Campbell S, Denery J, Aizawa S, Shibata S, Malahoff A, Alam M: **Genome sequence of the deep-sea gamma-proteobacterium Idiomarina loihiensis reveals amino acid fermentation as a source of carbon and energy.** *Proc Natl Acad Sci USA* 2004, **101**:18036-18041.
27. Jeong H, Yim JH, Lee C, Choi SH, Park YK, Yoon SH, Hur CG, Kang HY, Kim D, Lee HH, Park KH, Park SH, Park HS, Lee HK, Oh TK, Kim JF: **Genomic blueprint of Hahella chejuensis, a marine microbe producing an algicidal agent.** *Nucleic Acids Res* 2005, **33**:7066-7073.
28. Lauro FM, Bartlett DH: **Prokaryotic lifestyles in deep sea habitats.** *Extremophiles* 2007 in press.
29. Beatty JT, Overmann J, Lince MT, Manske AK, Lang AS, Blankenship RE, Van Dover CL, Martinson TA, Plumley FG: **An obligately photosynthetic bacterial anaerobe from a deep-sea hydrothermal vent.** *Proc Natl Acad Sci USA* 2005, **102**:9306-9310.

IV. MetaMine - A tool to detect and analyse gene patterns in their environmental context

Authors: Uta Bohnebeck, Thierry Lombardot, Renzo Kottmann, and Frank Oliver Glöckner

Published in *BMC Bioinformatics*. 2008; 9, no. 1 (October 28): 459.

Contribution: Use case definition, algorithm outline, system architecture, database

BMC Bioinformatics

BioMed Central

MetaMine - A tool to detect and analyse gene patterns in their environmental context

BMC Bioinformatics 2008, **9**:459 doi:10.1186/1471-2105-9-459

Uta Bohnebeck (bohnebeck@ttz-bremerhaven.de)
Thierry Lombardot (tlombard@mpi-bremen.de)
Renzo Kottmann (rkottman@mpi-bremen.de)
Frank O Glockner (fog@mpi-bremen.de)

ISSN 1471-2105

Article type Software

Submission date 25 April 2008

Acceptance date 28 October 2008

Publication date 28 October 2008

Article URL http://www.biomedcentral.com/1471-2105/9/459

MetaMine – A tool to detect and analyse gene patterns in their environmental context

Uta Bohnebeck[1§], Thierry Lombardot [2], Renzo Kottmann[2,3] and Frank O. Glöckner[2,3§]

[1]ttz Bremerhaven, An der Karlstadt 10, D-27568 Bremerhaven, Germany

[2]Max Planck Institute for Marine Microbiology, Celsiusstraße 1, D-28359 Bremen, Germany

[3]Jacobs University Bremen, Campus Ring 1, 28759 Bremen

[§]Corresponding authors

Email addresses:

UB: bohnebeck@ttz-bremerhaven.de

TL: tlombard@mpi-bremen.de

RK: rkottman@mpi-bremen.de

FOG: fog@mpi-bremen.de

Abstract

Background

Modern sequencing technologies allow rapid sequencing and bioinformatic analysis of genomes and metagenomes. With every new sequencing project a vast number of new proteins become available with many genes remaining functionally unclassified based on evidences from sequence similarities alone. Extending similarity searches with gene pattern approaches, defined as genes sharing a distinct genomic neighbourhood, have shown to significantly improve the number of functional assignments. Further functional evidences can be gained by correlating these gene patterns with prevailing environmental parameters. MetaMine was developed to approach the large pool of unclassified proteins by searching for recurrent gene patterns across habitats based on key genes.

Results

MetaMine is an interactive data mining tool which enables the detection of gene patterns in an environmental context. The gene pattern search starts with a user defined environmentally interesting key gene. With this gene a BLAST search is carried out against the Microbial Ecological Genomics DataBase (MEGDB) containing marine genomic and metagenomic sequences. This is followed by the determination of all neighbouring genes within a given distance and a search for functionally equivalent genes. In the final step a set of common genes present in a defined number of distinct genomes is determined. The gene patterns found are associated with their individual pattern instances describing gene order and directions. They are presented together with information about the sample and the habitat. MetaMine is implemented in Java and provided as a client/server application with a user-friendly graphical user interface. The system was evaluated with environmentally relevant genes related to the methane-cycle and carbon monoxide oxidation.

Conclusions

MetaMine offers a targeted, semi-automatic search for gene patterns based on expert input. The graphical user interface of MetaMine provides a user-friendly overview of the computed gene patterns for further inspection in an ecological context. Prevailing biological processes associated with a key gene can be used to infer new annotations and shape hypotheses to guide further analyses. The use-cases demonstrate that meaningful gene patterns can be quickly detected using MetaMine.
MetaMine is freely available for academic use from http://www.megx.net/metamine.

Background

More than 99% of the microbial diversity on earth still resists cultivation. To address their metabolic potential, numerous efforts to clone and sequence large DNA-fragments directly from the environment (the metagenome) have been accomplished worldwide. Several studies [1-3] have shown that on average only for 27-48% of the genes a specific function can be inferred by similarity searches [4]. This untapped pool of so called hypothetical proteins represents a still unexploited source for new enzymatically driven reactions.

With the availability of a large number of complete genomes comparative genomics becomes increasingly important. The major aspect of this approach is the analysis of gene neighbourhoods to indicate functional association, which can therefore significantly improve the predictions of gene functions. This idea was first systematically applied in 1999 by Overbeek and colleagues [5]. They introduced the concept of a "pair of close bidirectional best hits" and could prove functional coupling of such genes for several pathways. Some years later the subsystem approach was introduced [6, 7], which is a generalization of the pathway concept and describes a group of related functional roles jointly involved in a specific aspect of the cellular machinery. Today systems like IMG [8] / IMG/M [9], STRING [10, 11] and RAST [12] provide several functionalities to analyse gene neighbourhoods and other protein interaction features for a broad range of microbial genomes based on pre-computed data.

Additional support for functional evidences might be gained by correlating these gene patterns with prevailing environmental parameters. The growing number of genome and metagenome sequences from the environment, especially from the marine system, opens for the first time the possibility to link genomic information with environmental parameters in a systematic way [13-15]. Subsequently, if process correlated gene patterns can be identified in the habitats it should be possible to return hints on the functions of the respective patterns.

MetaMine is an interactive data mining tool which enables the detection of gene patterns, defined as genes sharing a distinct genomic neighbourhood, initiated by a user provided key gene. The underlying pipeline was designed to handle genomic data sets where gene family classification information is not available or incomplete. Because consistent family classification is a prerequisite for the pattern determination step we first calculate gene groups of functional equivalence. This is still an open research problem. Therefore, our system allows the user to work with different parameter settings and to switch between alternative methods. By focussing on the resulting patterns the calculation of the functional groups allows the inclusion of highly similar paralogs and genes to be part in several groups, because errors made in this step can be easily revealed by the gene patterns found. Given a user selected key gene of environmental relevance MetaMine carries out a semi-automatic search for gene patterns on a regularly updated database of marine genomes and metagenomes. The gene patterns found are presented to the user together with information about the samples and the habitat within an interactive graphical user interface (GUI) for further inspection. To our knowledge, currently no system exists combining genomic and metagenomic pattern information with environmental parameters.

Implementation

Microbial Ecological Genomics Database

The Microbial Ecological Genomics DataBase (MEGDB) contains prokaryotic genome and metagenome sequences of marine origin together with information about their environmental context. A list of criteria have been used to select the (meta-)genomic sequences to build the database: i) public access (sequences must be available in one of the public sequence databases [16]); ii) marine origin; iii) bacterial or archaeal; iv) high sequence quality (i.e. assembled contigs with a sequencing coverage of at least eight fold) and v) the exact geographic origin of the sequences (e.g. from the original publication).
Habitat parameters like water and sediment depths, temperature, salinity, and other physical-chemical properties have been extracted from the literature or extrapolated based on global ocean data sets like the World Ocean Atlas and the World Ocean Database [17] and remote sensing information (SeaWiFS) [18] within the EU project MetaFunctions [19]. A detailed overview of the current content of MEGDB can be found at http://www.megx.net/content.
Besides MetaMine, public access to the MEGDB is granted by the MetaLook tool [20] and especially the Genomes Mapserver as a central entry point [14, 21]. A Geographic-BLAST tool is also available online to get an overview of the geographical distribution of particular genes across the world.

Gene patterns and key gene approach

The term "gene pattern" often covers two related biological observations in genetics/genomics. In prokaryotic genomes, genes are often organised in operons, where transcription leads to a single messenger RNA molecule (mRNA) encoding different subunits of a protein or even distinct, but related proteins. The definition of an operon, as a set of commonly regulated genes is strict, but as long as no common mRNA is proven experimentally, the corresponding set of genes is often called a gene cluster/gene pattern which is loosely defined as a set of neighbouring genes with possibly coupled functions and/or conserved order across organisms.
The distinction between operons and gene patterns is not crucial for the biological questions we want to address with MetaMine. Moreover, the detection of habitat specific gene patterns requires some extensions with respect to environmental parameters in the concept described above. For the systematic search for correlations between the habitat and the gene content two basic types of gene patterns are of interest: (1) genes which are present or over/under-represented under specific environmental conditions and (2) patterns consisting of a set of genes occurring in specific genomic neighbourhoods. If such gene patterns are found genomic context analysis assists in potential functional assignments. In addition, if gene patterns correlate with distinct environmental parameters or processes further evidences for potential functions may be inferred. MetaMine was designed to detect such gene patterns. Due to the huge amount of genomic and metagenomic sequences we decided to apply a bottom-up approach, where prior biological knowledge is used to select a so-called *key gene* with known biological function and environmental relevance which plays the role of a seed to search for gene patterns in genomic sequences with at least two predicted genes.

Process steps

For gene pattern discovery the user can carry out the following process steps starting with the selection of the key gene. A detailed description of the analysis process including a flowchart can be found in the user guide available on the website and as Additional file 2.

1. & 2. Definition of a project and a key gene: In order to store and retrieve the results of a certain analysis the corresponding processing steps are organised in a project described by a project name, the user and a short comment. Also the key gene is defined by a name, a short description of its function and a comment.
3. Import/retrieval of the corresponding key gene sequence: The corresponding key gene sequence(s) can be imported from an external file containing the protein sequence in Fasta-format or by retrieving the protein sequence from the MEGDB.
4. BLAST search with key gene: Using the key gene sequence as a query for a BLAST search [22] against all marine genomic and metagenomic sequences with at least two predicted genes stored in the MEGDB is carried out. The result is a table with information about *e*-value, organism, sampling site, habitat, water and sediment depths and potential gene functions of similar genes found by this BLAST search presented to the user in a specific BLAST panel within the MetaMine GUI.
5. Determination of neighbouring genes: Given a user-defined parameter k the k neighbouring genes to each side with respect to all the genes found by the BLAST search in the previous step are determined and shown in tabular form. Using mouseover as well as a second panel the user can see detailed information about the functional annotation of the genes.
6. BLAST search of all neighbouring genes: A BLAST search is carried out with all neighbouring genes from the last step. The results are represented by a hash table containing the set of all neighbouring genes together with their associated BLAST results. The user can access this hash table by clicking a gene in the neighbour table and gets the associated BLAST information in the corresponding BLAST panel.
7. Determination of functionally equivalent genes: In order to detect functionally equivalent genes a reciprocal best hit approach followed by a clustering algorithm is carried out. The result is a set of groups, whereby each group represents genes of functional equivalence. All group members are colour-coded and presented to the user in the table of neighbouring genes (Fig. 1). In addition, the reciprocal best matches and the functional groups are shown in separate views.
8. Determination of gene patterns: Given two user-defined parameters minimal length l and quorum q a gene pattern is defined as a set of at least l genes (functional groups) which are all present in at least q different genomes or metagenome samples. Each pattern is associated with a pattern instance view. Pattern instances describing also gene order and directions are shown with their environmental information in tabular form (see also Fig 1). Guidelines for parameter settings can be found in the user guide (see Additional file 2).
9. Storage and retrieval of all intermediate results: All intermediate results are organised in special data objects which can be stored to and retrieved from the local MEGDB (stand-alone version only) as well as exported to and imported from external XML files.

Each process step can be repeated with other parameters resulting in a tree structure to organise intermediate results. As shown in the left panel of Fig. 1, the user can navigate through the history of all steps to analyse the corresponding results in detail.

The user should be aware that the final results can be influenced by the methods and parameter settings of all previous steps. Therefore, the differences can be used to prove the stability of the results. In case a certain gene is not in the functional group or gene pattern as expected this roll back mechanism allows further in depth analysis.
If the user specifies all parameters in advance he can also start a batch-mode analysis. All parameters can be adjusted using the Parameter Dialog "Set Parameters" in the Settings menu.

Algorithms

The following section gives a short overview of the underlying algorithms describing basic ideas and strategies.

Determination of functionally equivalent genes

In this step we are interested in finding groups of functionally equivalent genes which constitute the elements for the next step - the determination of common gene patterns. Different concepts and methods to obtain such groups exist. The classical and well established method - introduced by Clusters of Orthologous Groups (COG) [23] - relies on the phylogeny-based concept of orthology. Orthology describes genes in different species that have derived from a common ancestor by speciation in contrast to paralogous genes which arise from a duplication event. Therefore, orthology represents a strong relationship with a high potential describing the same biological function. Nevertheless, it is a phylogenetic concept originally introduced to study gene evolution. Consequently this excludes paralogs which might still have same functions. A complementary concept is to model intrinsic properties for gene function derived from multiple alignments and domain architecture as it is applied by Pfam [24, 25]. A third variant are automatic methods based on sequence similarity and unsupervised clustering like TRIBES [26].

A large proportion of our data set consists of metagenome samples with a high potential of new gene sequences not present in existing databases which might form new and sometimes small functional groups. Therefore, we started with the basic idea of COG and relaxed the constraints for metagenomes and inparalogs [27]. In general orthology is a well established concept for functional annotation and with the establishment of the gene patters the potential error of including some false positives is easily ruled out. In this context Boekhorst and Snel [28] have shown that "sharing gene order and similarity in size dramatically increases the chance of a query-hit pair being homologous".
To detect the groups of functionally equivalent genes we use a heuristic approach restricted to the gene sequences found in the BLAST searches which consist mainly of the following two steps:

- determination of reciprocal best matches and
- determination of groups of functionally equivalent genes.

Let G denote a set of identifiers for all genome and metagenome sequences which are stored in MEGDB and associated with an organism name and a sampling site, respectively and R denote a set of identifiers for all sequence regions which are predicted to be a protein encoding gene, then $genome$: $R \rightarrow G$ is a function determining the genome identifier for a certain gene. Let b_r denote the result of a BLAST search for gene $r \in R$ against the MEGDB and B a set of BLAST results b_r for a given set R,

then *rbm*: $R \times B \times G \rightarrow R$ is a function determining the identifier of the reciprocal best BLAST hit for gene $r \in R$ with respect to genome $g \in G$.
A reciprocal best match is commonly defined as follows: Gene g_a in genome G_A is the best match of gene g_b in genome G_B and gene g_b is the best match of gene g_a. Given the set of BLAST results B the function *rbm* checks this constraint for the genes and genomes specified. The search for reciprocal best matches is restricted to the set $R_n \subset R$ representing the genes found in the BLAST search with the key gene and their neighbours. Therefore, R_n corresponds to the upper table in the right panel in Fig. 1. In addition, the set of genomes $G_n \subset G$ is restricted to $\cup_{r \in Rn}$ *genome*(r). These are the genomes related to the genes (and their neighbours) found in the BLAST search with the key gene, which correspond to the rows of this table. The result of this step is a hash map *RBM* with key $r \in R_n$ storing a vector of genome related reciprocal best matches $rbm=(rbm_{g1}, rbm_{g2}, \ldots, rbm_{gn})$ with $\forall g_i \in G_n$ for each gene from the neighbour table. This intermediate result can be seen in the second table in the Orthology Panel.
The next step is the determination of functional groups which is based only on the information stored in the hash map *RBM* and carried out in a bottom-up manner. Let F denote this set of functional groups where each group $f_i \in F$, $i=1,\ldots,|F|$ contains a set of functionally equivalent genes, they are established as follows:

- For each gene $r \in R_n$ the corresponding vector of genome related reciprocal best matches is retrieved from the hash map *RBM* and checked for triangle relationships of reciprocal best matches. If such a triangle relationship exists between at least three genes of the vector a potential group f_{new} is created with these genes. This strategy corresponds to the COG approach [23].
- Check the new group against all already existing groups F for the following three cases: a) all genes of f_{new} are contained in a group f_i. Then f_{new} is not needed and will be removed. b) If the intersection of f_{new} with a group f_i is ≥ 3 genes and there are remaining genes in f_{new} not contained in group f_i, check these remaining genes for triangle relationships in f_i and include them if possible. If all remaining genes could be included in group f_i, f_{new} is not needed and will be deleted. c) There exists at least one gene from f_{new} which could not be included in any group f_i then f_{new} is added to the set F.
- Check all groups of set F for subset relationships. Delete the smaller one from F and keep only one set in case of equivalence.

Based on this procedure a gene can be part of several functional groups, a functional group can contain several genes from the same genome (inparalogs) but outparalogs are excluded by the *rbm* approach.

Determination of gene patterns
As described above, for our approach we define a gene pattern as a set of shared genes within a given genomic neighbourhood. This definition corresponds to a problem known as *gene team model* [29-32], which searches for a set of gene groups that co-occur in a given set of genomes. Further information on formal models can be found in chapter 8 of Mandoiu [33]. The order and the orientation of the genes need not be conserved, and insertions/deletions are allowed within the gene patterns. For in depth analysis we use the concept of pattern instance describing these properties, which are neglected in the process of pattern determination. The approaches mentioned above [29-32] are different with respect to the following features: i) if they are designed for two or more input genomes, ii) if they restrict a gene to be unique in

a genome/chromosome or if paralogs are allowed. In addition, these approaches require consistent family assignments of genes for all input genomes, which is not available or incomplete in many cases. Hu and colleagues [34] call this type of problem *gene pattern mining problem* and describe a very similar approach compared to ours.

For the pattern discovery step we have implemented two methods: a systematic search and a heuristic to reduce the search space. The systematic search is adapted from the character enumeration approach successfully applied in motif search algorithms like Pratt [35, 36] and TEIRESIAS [37] with the difference that the basic unit to enumerate is a functional group instead a single character.

Given the set of functional groups F of the previous step ordered by a numerical identifier and the parameters *minPatternLength* and q (quorum) describing the minimal length of a pattern as well as the minimal number of different sequences in which a pattern should be present, the systematic search is carried out as follows:

Let P denote the set of pattern to be determined, then P is initialized with patterns of length 1 represented by the entities $f \in F$. Each pattern is associated with a set of (meta-)genome identifiers $G_p \subset G$ where it occurs. In each iteration i, $i=2,...,|F|$ all patterns $p \in P \wedge |p|=i-1$ (the patterns from the last iteration having length i-1) are enlarged by the functional group having the next higher identifier and checked whether the corresponding set of genome identifiers covers more or equal entities than q. If yes, the pattern will be added to P. This systematic search guarantees to find all patterns fulfilling the constraints of the given parameters, but it has an exponential growing search space depending on the number of functional groups $|F|$: $O\left(\sum_{i=1}^{|F|}\binom{|F|}{i}\right)$

Therefore, a second method was implemented combining the systematic search and a heuristic. In order to generate a pattern there are two entities, which have to be checked: i) the functional groups as constituents for a pattern and ii) the (meta-) genome(s), where the pattern is present. In contrast to the systematic search based on the functional groups the heuristic inverts the constituents and the test in the following sense. First the gene patterns are generated as described above until a user-specified length *minLengthHeuristics* with default value of five. Second, for the set P of patterns found so far, the associated sets of (meta-)genome identifiers G_p are collected and filtered to be redundancy-free. The resulting set contains all genome combinations G_p, where a pattern can occur. Then, for each (meta-)genome combination G_p the largest set of shared functional groups are determined applying the intersection operation. Given a set of genome combinations the advantage of this heuristic is its ability to detect long gene patterns very quickly without explicit generation and testing of all functional group combinations, which can be a huge number.

System architecture

The system was implemented using a three-tier architecture that allows flexibility for subsequent integration of MetaMine into other systems. It can be used in two modes: as stand-alone system with direct access to a local database or as a client-server application using web services for all database operations.

The persistence layer is responsible for the permanent storage and retrieval of all necessary data for MetaMine. Therefore, it provides storage and retrieval functions for the MEGDB. In addition, there are functions to read and write to the file system especially for the import of key gene sequences stored in FASTA format and to import and export the analysis results as XML files for further data exchange. The

BLAST database file containing the protein sequences for the BLAST searches belongs also to this layer. In principle it is possible to exchange the underlying database with an own version.
The application layer contains all objects and methods implementing the application logic by providing methods for all functions provided to the user (see description of prototype for details). In addition, interfaces to external programs like BLAST for the sequence similarity search exist, as well as readers and writers for specific file formats used in molecular biological applications.
The presentation layer comprises the graphical user interface and the controller which activates the functions chosen by the user.

Results and discussion

The MetaMine software was tested using the MEGDB which contains high-quality georeferenced marine genomes and metagenomes of prokaryotic origin. This pattern detection approach can also successfully applied on large metagenome data sets based on shot-gun sequencing approaches as long as a significant fraction contains sequences with at least two predicted genes. Harrington *et al.* [4] reported recently that 47% of short metagenome sequences obtained by whole genome shot-gun sequencing in fact have neighbours even in the same transcriptional direction.

The focus of MetaMine is searching for gene patterns representing biological functions occurring in specific environmental contexts regardless their evolutionary history. Even if the current public DNA sequence databases covers only a fraction of the natural prokaryotic diversity [38], numerous environmentally relevant microbial pathways occurring in marine environments have been discovered and genetically described [39-41]. Two examples illustrating the benefits of the semi-automatic gene pattern discovery procedure of MetaMine for the study of globally important metabolic pathways in (meta-)genomic context are presented.

Archaea C1 metabolism gene patterns

Methanogenesis and the Anaerobic Oxidation of Methane (AOM) are two microbial metabolic pathways of environmental relevance, because they produce and consume the greenhouse gas methane in marine sediments, respectively. A MetaMine analysis using archaeal C1 key genes (mcrA, mcrB, mcrC, mcrD, mcrG, mrtC taken from [42]) discovered five distinct gene patterns called mcrB/G/A-14, mcrC-14, mcrB/D/C/G/A-5a/11, mcrC/B/G/A-5b and mrtC-17, where the name describes the key genes contained and the length of the patterns (see Fig. 2 and Additional file 1). As expected, the analysis shows that all key genes and their associated patterns occur exclusively in the habitat type “sediment”. The computational results foster the functional coupling of the genes with respect to their involvement in the C1 metabolism. Furthermore, all patterns, except that found in isolated organisms (mcrB/D/C/G/A-5a/11), revealed conserved hypothetical genes (chp; all red genes in Fig. 2) indicating their potential role in these particular metabolisms. These genes represent interesting new functional candidates and should be prior targets for wet-lab experiments. Moreover, four of the five gene patterns could be detected on metagenomic fragments, but not on complete genomes, which might reflect the specific modifications needed for the AOM metabolism as compared to the classical methanogenesis pathway [43, 42, 39].

Carbon monoxide oxidation gene patterns

Carbon monoxide (CO) is a gas evaporating from the ocean into the atmosphere. CO reacts with hydroxyl radicals who are also able to oxidize methane and nitrous oxides and is therefore an indirect mediator of the greenhouse effect [40]. Interestingly, microorganisms from the surface ocean water have recently been shown to carry genes encoding CO oxidation pathways, potentially influencing the diffusion of this gas in the atmosphere [44, 45].

In order to search for gene patterns associated with CO oxidation, the corresponding key genes have been used as input for a MetaMine analysis (coxL, GenBank: AAV95654 and GenBank: AAV94806 [44, 40]). The results show four main gene patterns including up to five genes (Fig. 3 and Additional file 1). Two conserved hypothetical genes can be found within these patterns, which designate them as potentially relevant for the CO oxidation pathways (Fig. 3, green and blue genes). Furthermore, one out of the eight gene patterns could be found exclusively in genomes isolated from marine sediments/geothermal sources, but not in genomes originating from the water column (Fig 3, pattern ID 70).

Conclusion

The exponentially growing DNA sequence datasets can only be handled effectively using semi-automatic processing pipelines that go beyond similarity based approaches. It has been shown that comparative approaches can significantly improve the quantity and quality of functional assignments leading to deeper insights into the complex metabolic and regulatory processes in a cell. The ecogenomic revolution initiated by the Venter cruises some years ago has opened the door to expand this approach by correlating syntenic gene patterns with specific environmental parameters and associated prevailing biological processes.

MetaMine offers a targeted, knowledge driven system to detect gene patterns for subsequent correlation with environmental information. First, the system is meant to confirm existing biological knowledge about genes involved in specific processes or pathways. Second, the approach has the potential to detect genes of so far unknown functions but functionally linked to specific habitat parameters. By integrating structural genomic information with environmental conditions MetaMine helps to find the "needle in the (meta)genomic haystack" especially for genes of so far unknown function. This reduced set of genes contains than prime candidates for further detailed functional analysis in the wet-lab. A use-case for *Archaea* C1 metabolism and CO oxidation genes showed that meaningful initial results can be quickly generated using MetaMine and a set of user-defined key genes. Further developments will concentrate on the incorporation of further genomic and metagenomic sequences, additional environmental parameters and further methods for the detection of functionally equivalent genes. In addition, to enhance the usability of MetaMine we plan to include links to external resources like GO or KEGG to support functional annotation as well as concepts to compare the functional groups found by MetaMine with other systems like COG.

Availability and requirements

Project name: MetaMine
Project home page: http://www.megx.net/metamine

Operating systems: Every OS with Java JRE 1.5 or higher (tested on Windows/ Linux).
Programming language: Java.
Other requirements: Java JRE 1.5 or higher
License: license-free.
Any restrictions to use by non-academics: MetaMine may not be sold or bundled with any type of commercial application.

Abbreviations

AOM: Anaerobic Oxidation of Methane; **CO**: carbon monoxide; **coxL**: carbon monoxide dehydrogenase (large subunit) gene; **mcr/Mcr**: methyl-coenzyme-M reductase gene/protein; **rbm**: reciprocal best match; **COG**: clusters of orthologous groups of genes; **GUI**: graphical user interface; **XML**: extensible markup language; **FP6**: the sixth framework programme of the European Union; **NEST**: new and emerging science and technology

Authors' contributions

UB designed and implemented MetaMine and drafted the manuscript. RK designed and implemented the current version of the underlying database and integrated the metagenomic data. UB and TL carried out and evaluated the biological test examples. UB, RK and TL participated in installing the MetaMine system as client server version at MPI. FOG is leading the EU-project MetaFunctions, gave advises for software development and has made revisions and contributions to the manuscript.

Acknowledgements

We thank the reviewers for their useful comments to improve the manuscript and all MetaFunctions partners for fruitful discussions. This work was supported by the FP6 EU project MetaFunctions (grant CT 511784), the Network of Excellence “Marine Genomics Europe” and the Max Planck Society.

References

1. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428**(6978):37–43.
2. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**(5667):66–74.
3. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative metagenomics of microbial communities.** *Science* 2005, **308**(5721):554–557.
4. Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, Raes J, Bork P: **Quantitative assessment of protein function prediction from metagenomics shotgun sequences**. *PNAS* 2007, **104**(35):13913–13918.
5. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proceedings of the National Acadademy of Sciences U S A* 1999, **96**(6):2896–2901.
6. Ye Y, Osterman A, Overbeek R, Godzik A: **Automatic detection of subsystem/pathway variants in genome analysis**. *Bioinformatics* 2005, **21 Suppl 1**:i478–i486.
7. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes**. *Nucleic Acids Research* 2005, **33**(17):5691–5702.
8. Markowitz V, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavromatis K, Ivanova N, Kyrpides N: **The integrated microbial genomes (IMG) system.** *Nucleic Acids Research* 2006, **34 (Database issue)**:D344–D348.
9. Markowitz V, Ivanova N, Palaniappan K, Korzeniewski ESF, Lykidis A, Anderson I, Mavromatis K, Kunin V, Martin HG, Dubchak I, Hugenholtz P, Kyrpides N: **An experimental metagenome data management and analy-sis system.** *Bioinformatics* 2006, **22**(14):e359–e367.
10. Snel B, Lehmann G, Bork P, Huynen MA: **STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene**. *Nucleic Acids Research* 2000, **28**(18):3442–3444.
11. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P: **STRING 7–recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Research* 2007, **35**(Database issue):D358–D362.
12. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD,

Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9**:75.

13. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Ashburner M, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, dePamphilis C, Edwards R, Faruque N, Feldman R, Glöckner FO, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kyrpides N, Leebens-Mack J, Lewis SE, Liste A, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Moxon R, Nelson K, Parkhill J, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, Wilson G, Wipat A: **Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification,** http://www.nature.com/nbt/consult/pdf/Field_et_al.pdf.
14. Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glöckner F: **Megx.net – database resources for marine ecological genomics.** *Nucleic Acids Research* 2006, **34 (Database issue)**:D390–D393.
15. Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IMA, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Ivanova NN, Kyrpides NC: **The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions.** *Nucleic Acids Research* 2008, **36**(Database issue):D528–D533.
16. **International Nucleotide Sequence Database Collaboration (INSDC)** http://www.insdc.org.
17. **National Oceanographic Data Center** http://www.nodc.noaa.gov.
18. **SeaWiFS Project** http://oceancolor.gsfc.nasa.gov/SeaWiFS/.
19. **EU project MetaFunctions** http://www.metafunctions.org.
20. Lombardot T, Kottmann R, Giuliani G, de Bono A, Addor N, Glöckner F: **MetaLook: a 3D visualisation software for marine ecological genomics.** *BMC Bioinformatics* 2007, **8**:406.
21. **Genomes Mapserver** http://www.megx.net/gms http://metafunctions.grid.unep.ch/mapserver.
22. Altschul S, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of Molecular Biology* 1990, **215**:403–410.
23. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution**. *Nucleic Acids Research* 2000, **28**:33–36.
24. Bateman A, Coin L, Durbin R, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E, Studholme D, Yeats C, Eddy S: **The Pfam protein families database**. *Nucleic Acids Research* 2004, **32 (Database issue)**:D138–D141.
25. Finn R, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, amd EL Sonnhammer SE, Bateman A: **Pfam: clans, web tools and services**. *Nucleic Acids Research* 2006, **34 (Database issue)**:D247–D251.
26. Enright A, Kunin V, Ouzounis C: **Protein families and TRIBES in genome sequence space**. *Nucleic Acids Research* 2003, **31**(15):4632–4638.

27. Alexeyenko A, Tamas I, Liu G, ELSonnhammer: **Automatic clustering of orthologs and inparalogs shared by multiple proteomes**. *Bioinformatics* 2006, **22**(14):e9–e15.
28. Boekhorst J, Snel B: **Identification of homologs in insignificant BLAST hits by exploiting extrinsic gene properties**. *BMC Bioinformatics*. 2007, **8**:356.
29. Luc N, Risler JL, Bergeron A, Raffinot M: **Gene teams: a new formalization of gene clusters for comparative genomics**. *Computational Biology and Chemistry* 2003, **27**:59–67.
30. Béal MPP, Bergeron A, Corteel S, Raffinot M: **An algorithmic view of gene teams**. *Theoretical Computer Science* 2004, **320**:395 – 418.
31. He X, Goldwasser MH: **Identifying Conserved Gene Clusters in the Presence of Homology Families**. *Journal of Computational Biology* 2005, **12**(6):638–656.
32. Kim S, JH, Yang CJ: **Gene teams with relaxed proximity constraint**. *Proc IEEE Comput Syst Bioinform Conf.* 2005:44–55.
33. Mandoiu I, (Eds) AZ: *Bioinformatics Algorithms: Techniques and Applications*. Wiley Book Series on Bioinformatics, John Wiley & Sons 2008.
34. Hu M, Choi K, Su W, Kim S, Yang J: **A gene pattern mining algorithm using interchangeable gene sets for prokaryotes**. *BMC Bioinformatics* 2008, **9**(124).
35. Jonassen I, Collins JF, Higgins DG: **Finding flexible patterns in unaligned protein sequences.** *Protein Science* 1995, **4**(8):1587–1595.
36. Jonassen I: **Efficient discovery of conserved patterns using a pattern graph.** *Computer Applications in the Biosciences* 1997, **13**(5):509–522.
37. Rigoutsos I, Floratos A: **Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm.** *Bioinformatics* 1998, **14**:55–67.
38. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families**. *PLoS Biol* 2007, **5**(3):e16.
39. Jørgensen B, Boetius A: **Feast and famine–microbial life in the deep-sea bed.** *Nature Reviews Microbiology* 2007, **5**(10):770–778.
40. Moran MA, Miller WL: **Resourceful heterotrophs make the most of light in the coastal ocean.** *Nature Reviews Microbiology* 2007, **5**(10):792–800.
41. Giovannoni S, Stingl U: **The importance of culturing bacterioplankton in the 'omics' age**. *Nature Reviews Microbiology* 2007, **5**(10):820–826.
42. Hallam S, Preston NPC, Detter J, Rokhsar D, Richardson P, DeLong E: **Reverse methanogenesis: testing the hypothesis with environmental genomics**. *Science* 2004, **305**(5689):1457–1462.
43. Meyerdierks A, Kube M, Lombardot T, Knittel K, Bauer M, Glöckner FO, Reinhardt R, Amann R: **Insights into the genomes of archaea mediating the anaerobic oxidation of methane**. *Environmental Microbiology* 2005, **7**(12):1937–1951.
44. Moran MA, Buchan A, González JM, Heidelberg JF, Whitman WB, Kiene RP, Henriksen JR, King GM, Belas R, Fuqua C, Brinkac L, Lewis M, Johri S, Weaver B, Pai G, Eisen JA, Rahe E, Sheldon WM, Ye W, Miller TR, Carlton J, Rasko DA, Paulsen IT, Ren Q, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Rosovitz MJ, Haft DH, Selengut J,

Ward N: **Genome sequence of Silicibacter pomeroyi reveals adaptations to the marine environment**. *Nature* 2004, **432**(7019):910–913.

45. Moran MA, Belas R, Schell MA, González JM, Sun F, Sun S, Binder BJ, Edmonds J, Ye W, Orcutt B, Howard EC, Meile C, Palefsky W, Goesmann A, Ren Q, Paulsen I, Ulrich LE, Thompson LS, Saunders E, Buchan A: **Ecological genomics of marine Roseobacters**. *Applied and Environmental Microbiology* 2007, **73**(14):4559–4569.

Figure legends

Figure 1 - MetaMine screenshot of the pattern panel

Figure 2 - Consensus patterns

Five consensus patterns (longest extension with mismatches) of the analysis with mcr gene. Chp represents conserved hypothetical proteins. All BLAST searches were carried out with a threshold of 1E-5.

Figure 3 - Analysis of coxL

All BLAST searches were carried out with threshold 1E-1.

Additional files

Additional file 1 – SupplementalMaterialPatterns.pps
The file contains screenshots of the analyses of the two examples and a more detailed description of the corresponding consensus patterns.

Additional file 2 – User_Guide_MetaMine1.2.pdf
The file contains a user guide about MetaMine version 1.2.

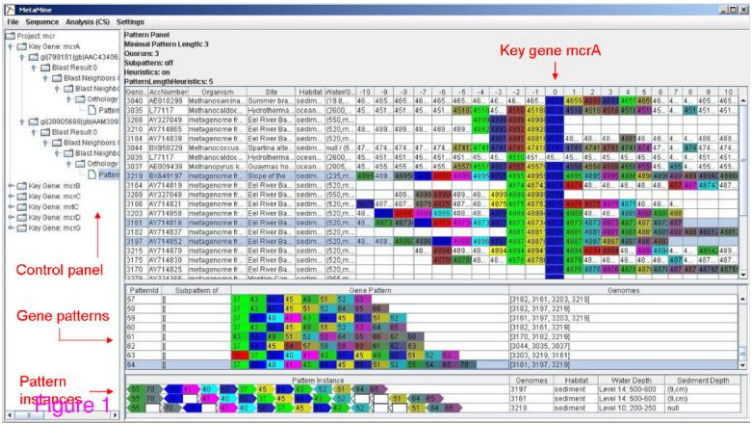

Figure 1

mcrC-14, found in metagenome Eel River Basin

chp	chp	-3	-2	-1	mcrC	chp	+2	+3	+4	chp	+6	chp	chp

-3: protease of the collagenase family; -2&-1: methyl coenzyme M reductase system component A2 (atw) +2: oxygen-sensitive ribonucleoside-triphosphate reductase; +3: pyruvate-formate lyase-activating enzyme +4: methyltransferase; +6: dGTP triphosphohydrolase

mcrC/B/G/A-5b, found in metagenome Eel River Basin

mcrC	hp	hp	chp	mcrB	mcrG	mcrA

mcrB/D/C/G/A-5a/11, found in Methanopyrus, Methanocaldococcus, Methanococcus, Methanosacrina, metagenome Eel River Basin

-1	mcrB	mcrD	mcrC	mcrG	mcrA	+1	+2	+3	+4	+5

-1: Fe-S oxidoreductase; +1: N5-methyl-tetrahydromethanopterin:coenzyme M methyltransferase, subunit E (mtrE); +2: mtrD; +3: mtrC; +3: mtrB; +4: mtrB; +5: mtrA

mrtC-17, found in metagenome Eel River Basin

-6	-5	-4	-3	-2	-1	mrtC	chp	chp	+3	+4	+5	+6	+7	+8	+9	+10

-6: Transcriptinal regulator; -5: LSU ribosomal proteins L19E; -4: L18p; -3: SSU ribosomal protein S5P; -2: L30p; -1: geranylgeranylglyceryl diphosphate synthase; +3: N5-methyl-tetrahydromethanopterin:coenzyme M methyltransferase, subunit E (mtrE); +4: mtrD; +5: mtrC; +6: mtrB; +7: mtrA; +8: mtrA; +9: mtrG; +10: mrtH

mcrB/G/A-14, found in metagenome Eel River Basin, Slope of North-Western Crimera area

chp	chp	-5	chp	-3	-2	-1	mcrB	mcrG	mcrA	chp	+2	chp	chp	+5	chp	+7	chp	+9	+10

-5: coenzyme PQQ synthesis protein (pqqE); -3: rhodanese-like thiosulfate sulfurtransferase; -2: protease of the collagenase family; -1: putative methyltransferase; +2: adenine deaminase (adeC); +5: pneumococcal surface protein; +7: multicatalytic endopeptidase complex subunit alpha; +9: exosome complex RNA-binding protein 1; +10: ribonuclease PH (rph)

Figure 2

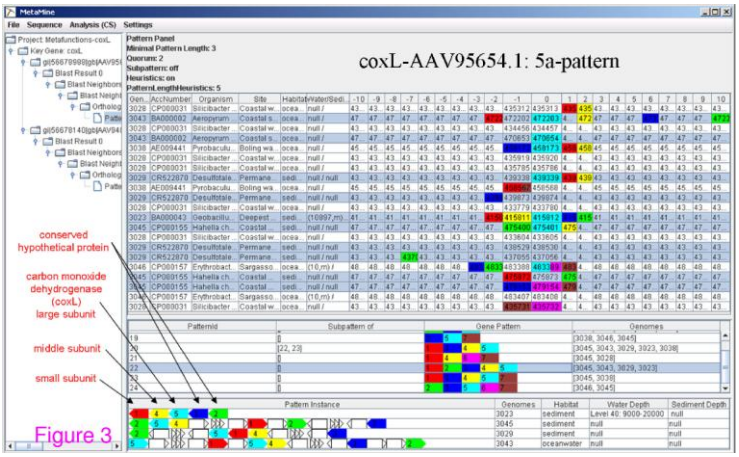

Figure 3

Additional files provided with this submission:

Additional file 1: supplementalmaterialpatterns_new.pps, 9634K
http://www.biomedcentral.com/imedia/7067738782115133/supp1.pps
Additional file 2: user_guide_metamine1.2-rev_final.doc, 532K
http://www.biomedcentral.com/imedia/8909720062208227/supp2.doc

v. The minimum information about a genome sequence (MIGS) specification.

Authors: Dawn Field, George Garrity, Tanya Gray, Norman Morrison, Jeremy Selengut, Peter Sterk, Tatiana Tatusova, et al.

Published in *Nature Biotechnology.* 2008; 26(5): 541-547.

Contribution: MIMS extension.

nature biotechnology

PERSPECTIVE

The minimum information about a genome sequence (MIGS) specification

Dawn Field*[1], George Garrity[2], Tanya Gray[1], Norman Morrison[3,4], Jeremy Selengut[5], Peter Sterk[6], Tatiana Tatusova[7], Nicholas Thomson[8], Michael J Allen[9], Samuel V Angiuoli[5,10], Michael Ashburner[11], Nelson Axelrod[5], Sandra Baldauf[12], Stuart Ballard[13], Jeffrey Boore[14], Guy Cochrane[6], James Cole[2], Peter Dawyndt[15], Paul De Vos[16,17], Claude dePamphilis[18], Robert Edwards[19,20], Nadeem Faruque[6], Robert Feldman[21], Jack Gilbert[9], Paul Gilna[22], Frank Oliver Glöckner[23], Philip Goldstein[24], Robert Guralnick[24], Dan Haft[5], David Hancock[3,4], Henning Hermjakob[6], Christiane Hertz-Fowler[8], Phil Hugenholtz[25], Ian Joint[9], Leonid Kagan[5], Matthew Kane[26], Jessie Kennedy[27], George Kowalchuk[28], Renzo Kottmann[23], Eugene Kolker[29–31], Saul Kravitz[5], Nikos Kyrpides[32], Jim Leebens-Mack[33], Suzanna E Lewis[34], Kelvin Li[5], Allyson L Lister[35,36], Phillip Lord[35], Natalia Maltsev[20], Victor Markowitz[37], Jennifer Martiny[38], Barbara Methe[5], Ilene Mizrachi[7], Richard Moxon[39], Karen Nelson[5,40], Julian Parkhill[8], Lita Proctor[26], Owen White[10], Susanna-Assunta Sansone[6], Andrew Spiers[42], Robert Stevens[3], Paul Swift[1], Chris Taylor[6], Yoshio Tateno[43], Adrian Tett[1], Sarah Turner[1], David Ussery[44], Bob Vaughan[6], Naomi Ward[45], Trish Whetzel[46], Ingio San Gil[41], Gareth Wilson[1] & Anil Wipat[35,36]

With the quantity of genomic data increasing at an exponential rate, it is imperative that these data be captured electronically, in a standard format. Standardization activities must proceed within the auspices of open-access and international working bodies. To tackle the issues surrounding the development of better descriptions of genomic investigations, we have formed the Genomic Standards Consortium (GSC). Here, we introduce the minimum information about a genome sequence (MIGS) specification with the intent of promoting participation in its development and discussing the resources that will be required to develop improved mechanisms of metadata capture and exchange. As part of its wider goals, the GSC also supports improving the 'transparency' of the information contained in existing genomic databases.

A wealth of genomic and metagenomic sequences

By the end of next year, there will be complete genome sequences of at least draft quality for more than 1,000 bacteria and archaea and 100 eukaryotes[1,2] and for even larger numbers of viruses, organelles and plasmids. With the rapid pace at which new genome sequences are appearing, the need to consider how best to ensure stewardship of these data for the long term has never been more pressing.

Our genome collection: more than the sum of its parts. The analysis of genomic information is having an impact on every area of the life sciences and beyond. A genome sequence is a prerequisite to understanding the molecular basis of phenotype, how it evolves over time and how we can manipulate it to provide new solutions to critical problems. Such solutions include therapies and cures for disease, industrial products, approaches for biodegradation of xenobiotic compounds and renewable energy sources. With improvements in sequencing technologies, the growing interest in metagenomic approaches and the proven power of comparative analysis of groups of related genomes, we can envision the day when it will be commonplace to sequence tens to hundreds of genomes or more as part of a single study. At current rates of genome sequencing, it has been estimated that >4,000 bacterial genomes will be available soon after 2010 (ref. 1).

Given the importance of the growing genome collection, the capital investment in its creation and the benefits of leveraging its value through diverse comparative analyses, every effort should be made to describe it as accurately and comprehensively as possible. There is an increasing interest from the community in doing so, for three main reasons. The first is the interest in testing hypotheses about the features observed in genomes using comparative evo- and eco-genomic approaches[3]. The second is the need to supplement the content of a variety of databases with high-level descriptions of genomes that allow useful grouping, sorting and searching of the underlying data. The third is the growth in genome sequence data from environmental isolates and metagenomes—vast data sets of DNA fragments from environmental samples[4–6]. The data generated by such studies will dwarf current stores of genomic information, making improved descriptions of genomes even more important.

At present, both top-level descriptors and genome descriptions are incomplete for many reasons. First and foremost, in hindsight we now know the minimum quality and quantity of information that is required to make each description precise, accurate and useful. For example, even for bacterial and archaeal species with validly published names, strain names were not routinely captured in genome annotation documents before the sequencing of large numbers of genomes from the same species[7],

*A list of affiliations appears at the end of the paper.

Published online 8 May 2008; doi:10.1038/1360

but such information is now considered essential. Through empirical observations, we are expanding our view of the types of information that are important for testing particular hypotheses, exploring new patterns and quantifying inherent sampling biases[3,8].

As the number of habitats and communities sampled using metagenomic approaches increases, we are also being forced to rethink our understanding of the minimum information required to adequately describe a genome sequence. Without adequate description of the environmental context and the experimental methods used, such data sets will be of less value for researchers wishing to conduct comparative genomic studies or link genetic potential with the diversity and abundance of organisms. In fact, given the vast number of uncultivated microbes, it may be that a DNA-centric approach, in which genes are linked to habitats (locations), is more useful than the species-centric view[9,10]. Finally, sequencing technology is advancing rapidly, and the adoption of new methods[11,13] will force the adoption of additional descriptors (e.g., the depth of sequence coverage, quality and whether any 'finishing' was used) to be able to distinguish among these methods.

Box 1 Minimum Information about a Genome Sequence (MIGS) checklist version 2.0

Investigation	Report type EU	BA	PL	VI	OR	ME
• Submit to trace archives and INSDC	M	M	M	M	M	M
• Investigation type (i.e., report type)	M	M	M	M	M	M
• Project name[2]	M	M	M	M	M	M
• Study						
• Environment						
• Geographic location (latitude and longitude[float (point, transect and region)], depth and altitude of sample)[(integer)]	M	M	M	M	M	M
• Time of sample collection[(UCT)]	M	M	M	M	M	M
• Habitat[EnvO]	M	M	M	M	M	M
MIMS extension: select to report a set of uniform measurements for a given habitat:						M
• Water body: (temperature, pH, salinity, pressure, chlorophyll, conductivity, light intensity, dissolved organic carbon (DOC), current, atmospheric data, density, alkalinity, dissolved oxygen, particulate organic carbon (POC), phosphate, nitrate, sulfates, sulfides, primary production)[(integer, unit)]						
• Nucleic acid sequence source						
• Subspecific genetic lineage (below lowest rank of NCBI taxonomy, which is subspecies) (e.g., serovar, biotype, ecotype)[(CABRI)]	M	M	M	M	M	–
• Ploidy (e.g., allopolyploid, polyploid)[(PATO)]	M					
• Number of replicons (EU, BA: chromosomes (haploid count); VI: segments)[(integer)]	M	M	–	M	–	–
• Extrachromosomal elements[(integer)]	X	M				
• Estimated size (before sequencing; to apply to all draft genomes)[(integer; base pairs)]	M	X	X	X	X	–
• Reference for biomaterial (primary publication if isolated before genome publication; otherwise, primary genome report)[(PMID or DOI)]	X	M	X	X	X	X
• Source material identifiers: (cultures of microorganisms: identifiers[(alphanumeric)] for two culture collections[(OBI)]; specimens (e.g., organelles and Eukarya): voucher condition and location[(CV)])	M	M	M	M	M	M
• Known pathogenicity		M		M		
• Biotic relationship (e.g., free-living, parasite, commensal, symbiont)[(OBI)]	X	M		X		
• Specific host (e.g., host taxid, unknown, environmental)[EnvO]	X	M	M	M		
• Host specificity or range[(taxid)]	X	X	X	M		
• Health or disease status of specific host at time of collection (e.g., alive, asymptomatic)[PATO]		M		M		
• Trophic level (e.g., autotroph, heterotroph)[PATO]	M	M	–	–	–	–
• Propagation (phage: lytic or lysogenic; plasmid: incompatibility group)[(CV)]	M		M	M	–	–
• Encoded traits (e.g., plasmid: antibiotic resistance; phage: converting genes)[(CV; see caption)]		X	M	M		X
• Relationship to oxygen (e.g., aerobic, anaerobic)[PATO]		M	–	–	–	–
• Isolation and growth conditions[(PMID or DOI)]	M	M	M	M	M	M
• Biomaterial treatment (e.g., filtering of sea water)[(OBI)]						M
• Volume of sample[(integer)]						M
• Sampling strategy (enriched, screened, normalized)[(CV)]						M

continued

Box 1 Minimum Information about a Genome Sequence (MIGS) checklist version 2.0 (continued)

• Assay						
• Sequencing						
• Nucleic acid preparation (extraction method(CV); amplification(CV)	M	M	M	M	M	M
• Library construction (library size(integer), number of reads sequenced(integer), vector(CV)						M
• Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony)(OBI)	M	M	M	M	M	M
• Assembly (assembly method(CV), estimated error rate(unit) and method of calculation(CV))	M	M	M	M	M	M
• Finishing strategy (status—e.g., complete or draft(CV), coverage(integer), contigs(integer))	M	M	X	X	X	X
• Relevant Standard Operating Procedures (SOPs)	M	M	M	M	M	M
• Relevant electronic resources	M	M	M	M	M	M

All proposed descriptors in MIGS and the reports (groups) to which they apply are listed. EU, eukaryotes; BA, bacteria and archaea; PL, plasmid; VI, virus; OR, organelle; ME, metagenome. Each descriptor has superscripts denoting its 'type' (e.g., integer or controlled vocabulary (CV) term). For items marked "CV," candidate OBO ontologies (http://obofoundry.org), if available, have been selected for use. EnvO, The Environment Ontology; PATO, the Phenotype and Trait Ontology; CABRI, Common Access to Biological Resources and Information. Mixed ontologies may be useful for the "encoded traits" descriptor: the PATO term "resistant" could be used with a ChEBI term—for example, "penicillin"—to note antibiotic resistance to a given compound. Descriptors in shaded rows are common to all report types and are considered the 'core' of MIGS. "Source material identifier" is an exception; the GSC recommends this be a core descriptor, but as yet, physical archives are not yet routinely created for all cases or types of biological material subjected to genome sequencing (the recommended deposition in at least two culture collections for viable samples[20] and vouchers for specimens). This is due to both cultural and technical issues. The need for universal and unique identifiers for metagenomic samples is an idea recently discussed in an exploratory workshop organized by the MetaFunctions group (http://www.metafunctions.org). In fact, the application of MIGS to our complete genome collection will require the designation of permanent and unique identifiers for all genome projects, something the INSDC is working to implement[21]. Geographic location is applied in principle to all report types, but we recognize that many isolates, especially eukaryotes, are highly domesticated laboratory organisms distantly separated from an environmental context of relevance. All descriptors deemed to be core are marked "M" (minimum) and others which could be optionally applied to other groups with high priority are marked "X" (extra). Taxonomic groups for which a descriptor cannot be meaningfully applied are marked with a dash. This list of minimal information is recognized by the GSC as just a starting point for the description of genomes and metagenomes. PMID, PubMed identifier; DOI, digital object identifier; float, floating-point decimal; UCT, Coordinated Universal Time (YYYY-MM-DD); unit, a suitable unit of measure. The descriptors isolation and growth conditions take citations as their values because the information can not be contained in a single value (or small set of values) like those of all other fields. This could be given as the PMID or DOI of the publication. It could also be an SOP. In principle, all aspects of the checklist could be substantiated with a reference in addition to a value, and this would be captured at the level of implementation.

Most often, metadata about genome sequences are found only in the primary literature or in reference works, such as *Bergey's Manual*[14] for bacteria and archaea, rather than in sequence databases. The distributed and patchy nature of this information and the difficulties of curating even a few pieces of information for what are now very large collections of genomes make the vision of a single definitive source of rich genomic descriptions highly desirable.

The need for coordinated efforts

Facilitating and accelerating the process of collecting relevant metadata would clearly reduce ongoing replication of efforts and maximize the ability to share and integrate data within the genomics community. The obvious solution is to develop a consensus-based approach.

The Genomic Standards Consortium. The GSC is an open-membership, international working body formed in September 2005 (ref. 15). Its goal is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data. The GSC community brings together (i) evolutionists, ecologists, molecular biologists and other researchers analyzing collections of genomes, (ii) bioinformaticians producing genomic databases, (iii) those who sequence genomes and (iv) computer scientists, ontology experts and members of other standardization initiatives, such as the International Nucleotide Sequence Database Collaboration (INSDC), which is responsible for the DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL) and GenBank databases (http://www.insdc.org/). The guidance of DDBJ, EMBL and GenBank will be critical to the success of the GSC initiative, both because they are the official stewards of the public collection of genomes and because of their interest in fulfilling community needs.

Minimum information about genomes and metagenomes

The GSC is working to define a set of core descriptors for genomes and metagenomes in the form of a MIGS specification (**Fig. 1**). MIGS extends the minimum information already captured by the INSDC. The MIGS checklist is given in **Box 1**, and the most up-to-date version is available from the consortium's website (http://gensc.sf.net). Examples of MIGS-compliant reports are given in **Supplementary Table 1** online. The information required to comply with MIGS is routinely included in primary genome publications (or is referenced therein). However, this information needs to be formalized and made available in electronic form to improve its accessibility[16] (**Box 2**).

Since it was originally proposed[16], the MIGS specification has been simplified and changed by the GSC through an iterative revision process to contain (i) only curated information that cannot be calculated from raw genomic sequence and (ii) core descriptors specific to the major taxonomic groups (eukaryotes, bacteria and archaea[17], plasmids, viruses, organelles) and metagenomes. MIGS is structured as an 'Investigation' composed of a 'Study' and an 'Assay', according to the Reporting Structures for Biological Investigations (RSBI) working group's recommendation for the modularization of checklists[18,19]. Under 'Study' are the top-level concepts 'Environment' and 'Nucleic Acid Sequence' and under 'Assay' is a description of the sequencing technology.

MIGS aims to support unencumbered access to genomic reagents (such as strains)[20], place the complete (meta)genome collection into geospatial and temporal context (latitude, longitude, altitude or depth, date and time of sampling) and provide essential details of the experimental method used (e.g., sequencing method). MIGS also provides a framework for the capture of extra information deemed 'minimum' to specific communities. Most importantly, the description of metagenomes in MIGS is being extended in the minimum information about a metagenome sequence (MIMS) specification[21]. MIMS enables the capture of further measurements that define habitat (such as temperature, salinity, pH, dissolved organic carbon) and extends the original structure of MIGS for describing a single (meta)genomic experiment to allow the capture of information from pooled samples and more than one independent sampling event (e.g., sampling along a transect[4]).

How genomes and metagenomes are described in public databases has evolved from how short, simple DNA sequences are described, without special attention to information such as the geographical origin of the sequence. Significant efforts are underway by the INSDC to adapt and extend the infrastructure for describing genomes through the Genome

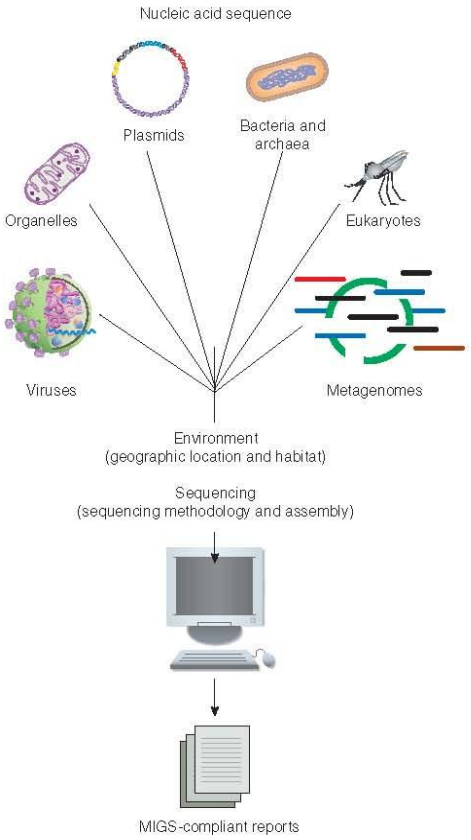

Figure 1 The scope of MIGS. The MIGS specification enables description of the complete range of possible genomes (eukaryotes, bacteria, archaea, plasmids, viruses, organelles) and metagenomes. Core descriptors include information about the origins of the nucleic acid sequence (genome), its environment (latitude and longitude, date and time of sampling and habitat) and sequence processing (sequencing and assembly methods). MIGS-compliant reports can be rendered into an electronic format using the MIGS XML schema and controlled vocabularies through the GSC's Genome Catalogue (http://gensc.sf.net).

Project Metadata initiative[22]. The INSDC efforts are open to evolution, albeit at a conservative pace[22], and it is the GSC's hope that much, if not all, of the MIGS specification will be included in the Genome Project Metadata initiative. A mapping between INSDC features and MIGS has been developed for the purpose of placing MIGS information into INSDC documents and is available on our website. Any fields that are not already formally defined by the INSDC Feature Table Document (http://www.insdc.org/files/documents/feature_table.html) can be represented within a structured comment block in INSDC records[22].

A genome catalog

The development of any checklist must be an open and iterative process that involves a balanced group of participants. Moreover, mechanisms for achieving compliance are needed to facilitate widespread adoption of a checklist. Such mechanisms involve an appropriate reporting structure for capturing and exchanging data (file formats), software, databases and appropriate controlled vocabularies and/or ontologies for defining the terms used in the annotations. The GSC is working toward these combined goals and has created an online system for capturing MIGS-compliant reports (http://gensc.sf.net).

In brief, we have implemented the checklist as an XML schema and built a freely available Genome Catalogue system (GCat) (http://gensc.sf.net). GCat is designed to generate forms automatically and 'on the fly' from this schema for the sake of data input. It also allows users to view and search genome descriptions as they accumulate during the process of refining the MIGS checklist. The GCat system is generic and could be applied to the capture of more expressive metadata for subsets of genomes. Indeed, it is flexible enough to support the implementation of any checklist that can be structured as an appropriate XML schema (MIGS.xsd, being developed into the Genomic Contextual Data Markup Language (GCDML)). The GSC is also working in the area of controlled vocabulary and ontology development through the collation of controlled vocabularies already in use in the community and through contributions to the Ontology for Biomedical Investigations (OBI, previously known as the Functional Genomics Investigation Ontology (FuGO)[23]) and the Environment Ontology (EnvO) project (http://environmentontology.org). As a part of this process, GCat makes use of existing controlled vocabulary terms and accepts new terms.

Improving genomic databases

By design, MIGS contains only primary, curated information. This is because secondary, or derived, information that can be calculated from a genome sequence is subject to frequent change, can be generated using more than one method and should be acquired directly from those producing the calculations. Still, access to computed information (e.g., in the simplest cases, G+C content or total number of predicted proteins) should be made as easy as possible.

Genomic sequences and their initial annotations must be submitted to the INSDC (http://www.insdc.org/) (and subsequent high-quality, curated annotations derived from empirical observations to the Third Party Annotation data set[24]), but there are an ever increasing number of genomic databases containing a wide range of additional computations. Although GSC does not endorse any particular method of analysis or database, it supports increased transparency of such resources for the sake of accurate data interpretation and integration.

The first issue is that of exchanging calculated information. This could be facilitated in part by widespread adoption of a common exchange format, such as the Generic Feature Format Version 3 (GFF3) file format (http://song.sourceforge.net/gff3.shtml). There are many tools that support the reformatting of a variety of file types into GFF3, so database providers would find it straightforward to generate appropriate files. The availability of a wide suite of tools for downstream analyses of files in GFF3 format also means that users could combine the weight of evidence from many sources when examining a particular genome. This could reveal instances of systemic bias and therefore lead to better genomic annotations, as more composite features would be available and conflicting annotations could be highlighted for resolution.

Exchanging data also relies on common standards for computational analyses, and supporting data downloads is not enough, regardless of format. Data resources should also be expected, within reason, to provide clear specifications for how the data are generated (for example, standard operating procedures (SOPs) that describe computations such as gene prediction and operon and ortholog identification). One example of this

type of documentation is provided in *AboutIMG*, a web-based description of the Integrated Microbial Genomes (IMG) system[25].

In the future it should be far simpler to combine various genomic features, exact details of how they were generated and enough information about the provenance (origin) of the analyses to be able to transparently share data from different sources. Such interoperability, especially when provided by participating databases in a way that would enable automatic harvesting of the data (e.g., through web service technology), would multiply the individual value of these databases many times over and open up new opportunities to examine genome sequences in unprecedented detail.

Future directions

The effort required to achieve the degree of transparency advocated here is considerable but offers substantial and immediate benefits. We argue that the cost of achieving such standardization is trivial compared with the sums spent generating the data. The capture of MIGS-compliant information will not only facilitate comparative genomic and metagenomic analyses but also enhance the available descriptions of downstream '-omic' experiments based on genomic data. It will also enhance the much larger 'halos' of 16S ribosomal RNA sequences that are now available for many sequenced genomes and metagenomes. For example, the genome sequence of the marine bacterium *Silicibacter pomeroyi*[26] is 'embedded' in a large number of environmental 16S rRNA sequences affiliated with the Roseobacter lineage, which is accompanied by a fairly extensive literature describing the distribution, ecology and other properties of this group[27].

Through its ongoing efforts, the GSC hopes to stimulate discussion of the MIGS specification and solicit further feedback from the community. It therefore has an open call for participation and is eager to

Box 2 Frequently asked questions about MIGS

Below we answer general questions about MIGS, its development and how to use it.

What is MIGS?

- MIGS specifies a formal way to describe genomes and metagenomes in more detail than is captured at present in DDBJ, EMBL and GenBank documents.
- The information in MIGS is intended to be used in comparative genomic analysis, provide a better understanding of the source of each genome and enable us to situate genomes and metagenomes in their geospatial and temporal contexts (when relevant) through the specification of geographic location and sampling date.

Do all genomes and metagenomes fall under the scope of MIGS?

- Yes. MIGS has elements describing eukaryotic, bacterial and archaeal, plasmid, viral and organellar genomes as well as metagenomes. Some of the core elements overlap between types of records, and some are unique to one or more groups.

Who has driven the development of MIGS?

- MIGS has been developed through a series of GSC workshops involving participants from DDBJ, EMBL, the US National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI), Joint Genome Institute (JGI), Sanger Institute, J. Craig Venter Institute (JCVI, formerly TIGR), Max Planck Institute, the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) project and a variety of other research institutions.

Who should complete a MIGS report?

- Authors of genome and metagenome publications should submit a report after submitting project information to DDBJ, EMBL or Genbank.

Is MIGS very time-consuming to complete?

- MIGS is a short specification compared with most other '-omic' checklists (see http://mibbi.sf.net) for three reasons:
- MIGS is an extension of the data already captured by DBBJ, EMBL and Genbank to describe genomes and metagenomes and is designed to be complementary to these authoritative sources of metadata. The INSDC genome project database will contain essential administrative information, taxonomy identifiers (taxids) and a genome project identifier (PID).
- MIGS was intentionally designed to be 'minimal' to encourage its adoption.
- Genomic sequences, unlike transcriptomes, proteomes or metabolomes, are 'state independent' (a genome sequence is stable with respect to cellular state and environmental factors). In contrast, metagenomic experiments depend on the sampling strategy and the specific habitat of a given microbial community, requiring a further specification (MIMS) to define habitat parameters such as salinity, pH and temperature.

How can I get a unique identifier for my submission for use in my publication?

- The Genomes Online Database (http://www.genomesonline.org) is the recognized authority for issuing GCat identifiers for eukaryotes, bacteria and archaea and metagenomes. The Genome Catalogue (GCat) will issue identifiers for other genomes.

Can I submit MIGS-compliant information online?

- Yes. The GSC has developed a portal called the 'Genome Catalogue' that has been useful in prototyping the MIGS specification. MIGS-compliant information can be submitted through user-friendly web forms with drop-down menus for the selection of appropriate terms; batch uploading functions are being developed (http://gensc.sf.net).

Are sample reports available?

- Yes, the Genome Catalogue contains a collection of MIGS-compliant reports. Examples are given in **Supplementary Table 1**.

How would I report the existence of MIGS-compliant data in my publication?

- MIGS-compliant information could be reported as a supplementary table in a publication. Far more beneficial to the wider community would be to submit this information to the Genome Catalogue and report the GCat identifier and the URL of this database.

How can I get involved in the GSC and provide feedback for the development of MIGS?

- The GSC has an open call for participation. Further information can be found at http://gensc.sf.net.

solicit MIGS-compliant genome reports (including batch uploads) and collect relevant controlled vocabulary terms useful in the description of genomes and metagenomes. GCat identifiers have been implemented and are available for past or future projects, and MIGS-compliant genome reports are starting to become available online (e.g., refs. 28–31). We expect a production version of MIGS (2.0) to be released by early 2008 with an appropriate set of terms formalized within OBI[19] and other relevant Open Biomedical Ontology (http://obofoundry.org/) ontologies. We would hope that this milestone (release of MIGS 2.0) will be accompanied by recognition by journals and implementation by a variety of databases. Beyond this, the MIGS specification should still remain flexible enough to allow it to be revised in accordance with advances in technology and our biological knowledge. It should also be considered for use in combination with other checklists in the context of the Minimum Information about a Biomedical or Biological Investigation (MIBBI) Foundry (http://mibbi.sf.net), of which the GSC is a founding community[19]. The most up-to-date information about GSC activities is available at our website (http://gensc.sf.net).

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

ACKNOWLEDGMENTS

We would like to thank the UK National Institute of Environmental eScience (NIEeS) and the European Bioinformatics Institute (EBI) for hosting GSC workshops and the UK Natural Environmental Research Council for providing funds for coordination (NE/D01252X/1) and infrastructure building activities (NE/E007325/1).

DISCLAIMER

Opinions, findings and conclusions or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the US National Science Foundation.

Published online at http://www.nature.com/naturebiotechnology
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
2. Liolios, K., Mavromatis, K., Tavernarakis, N. & Kyrpides, N.C. 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 (2008).
3. Martiny, J. & Field, D. Ecological perspectives on our complete genome collection. *Ecology Letters* **8**, 1334–1345 (2005).
4. Rusch, D.B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol. [online]* **5**, e77 (2007).
5. Edwards, R.A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics* **7**, 57 (2006).
6. Committee on Metagenomics: Challenges and Functional Applications, National Research Council. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (National Academies Press, Washington, DC, 2007).
7. Coenye, T. & Vandamme, P. Bacterial whole-genome sequences: minimal information and strain availability. *Microbiology* **150**, 2017–2018 (2004).
8. Haft, D.H., Selengut, J.D., Brinkac, L.M., Zafar, N. & White, O. Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* **21**, 293–306 (2005).
9. Lombardot, T. *et al.* Megx.net—database resources for marine ecological genomics. *Nucleic Acids Res.* **34** (database issue), D390–D393 (2006).
10. Tautz, D., Arctander, P., Minelli, A., Thomas, E. & Vogler, A.P. A plea for DNA taxonomy. *Trends Ecol. Evol.* **18**, 70–74 (2003).
11. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).
12. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
13. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
14. Garrity, G.M. (ed.) *Bergey's Manual of Systematic Bacteriology*, 2nd edn., Vol. 1, (Springer, New York, 2001).
15. Field, D. *et al.* Meeting report: eGenomics: cataloguing our complete genome collection I. *Comp. Funct. Genomics* **6**, 357–362 (2006).
16. Field, D. & Hughes, J. Cataloguing our current genome collection. *Microbiology* **151**, 1016–1019 (2005).
17. Pace, N.R. Time for a change. *Nature* **441**, 289 (2006).
18. Sansone, S.A. *et al.* A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. *OMICS* **10**, 164–171 (2006).
19. Taylor, C. *et. al.* Promoting coherent minimum reporting requirements for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* (in the press).
20. Ward, N., Eisen, J., Fraser, C. & Stackebrandt, E. Sequenced strains must be saved from extinction. *Nature* **414**, 148 (2001).
21. Field, D. *et al.* Meeting report: eGenomics: cataloguing our complete genome collection III. *Comp. Funct. Genomics* **2007**, 47304 (2007).
22. Morrison, N. *et al.* Concept of sample in OMICS technology. *OMICS* **10**, 127–137 (2006).
23. Whetzel, P.L. *et al.* Development of FuGO: an ontology for functional genomics investigations. *OMICS* **10**, 199–204 (2006).
24. Cochrane, G. *et al.* Evidence standards in experimental and inferential INSDC Third Party Annotation data. *OMICS* **10**, 105–113 (2006).
25. Markowitz, V.M. *et al.* IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* **36** (database issue), D534–D538 (2008).
26. Moran, M.A. *et al.* Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432**, 910–913 (2004).
27. Buchan, A., Gonzalez, J.M. & Moran, M.A. Overview of the marine roseobacter lineage. *Appl. Environ. Microbiol.* **71**, 5665–5677 (2005).
28. Angly, F.E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368 (2006).
29. Bauer, M. *et al.* Whole genome analysis of the marine Bacteroidetes '*Gramella forsetii*' reveals adaptations to degradation of polymeric organic matter. *Environ. Microbiol.* **8**, 2201–2213 (2006).
30. Glockner, F.O. *et al.* Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc. Natl. Acad. Sci. USA* **100**, 8298–8303 (2003).
31. Rabus, R. *et al.* The genome of *Desulfotalea psychrophila*, a sulfate-reducing bacterium from permanently cold Arctic sediments. *Environ. Microbiol.* **6**, 887–902 (2004).
32. Raes J., Foerstner, K.U. & Bork, P. Get the most out of your metagenome: com-

[1]Natural Environmental Research Council Centre for Ecology and Hydrology, Oxford OX1 3SR, UK. [2]Michigan State University, East Lansing, Michigan 48824, USA. [3]School of Computer Science, University of Manchester, Manchester M13 9PL, UK. [4]NERC Environmental Bioinformatics Centre, Oxford Centre for Ecology and Hydrology, Oxford OX1 3SR, UK. [5]J. Craig Venter Institute (JCVI), 9704 Medical Center Drive, Rockville, Maryland 20850, USA. [6]European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [7]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. [8]Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [9]Plymouth Marine Laboratory, Prospect Place, Plymouth PL1 3DH, UK. [10]Institute for Genome Sciences and Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, 20 Penn Street, Baltimore, Maryland 21201, USA. [11]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK. [12]Department of Biology, University of York Box 373, York, YO10 5YW, UK. [13]National Institute of Environmental eScience, Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK. [14]US Department of Energy (DOE) Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [15]Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium. [16]Laboratory of Microbiology, Ghent University, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium. [17]BCCM/LMG Bacteria Collection, Ghent University, K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium. [18]Penn State University, 208 Mueller Laboratory, University Park, Pennsylvania 16802, USA. [19]Department of Computer Science, 5500 Campanile Drive, San Diego State University, San Diego, California 92182, USA. [20]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. [21]SymBio Corporation, 1455 Adams Drive, Menlo Park, California 94025, USA. [22]California Institute for Telecommunications and Information Technology (Calit2), a University of California San Diego (UCSD)/University of California Irvine partnership, 9500 Gilman Drive, La Jolla, California, 92093, USA. [23]Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, Bremen 28359 Germany. [24]Department of Ecology and Evolutionary Biology and University of Colorado Natural History Museum, 218 UCB, University of Colorado, Boulder, Colorado 80309, USA. [25]Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Building 400-404, Walnut Creek, California 94598, USA. [26]The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA. [27]School of Computing, Napier University, Merchiston Campus, 10 Colington Road Edinburgh, Scotland, EH10 5DT, UK. [28]Department of Terrestrial Microbial Ecology, Netherlands Institute of Ecology, Centre for Terrestrial Ecology, PO Box 40, Heteren 6666 ZG, Netherlands. [29]BIATECH Institute, 19310 North Creek Parkway South, Suite 115, Bothell, Washington 98011, USA. [30]Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Information, University of Washington, Seattle, Washington 91895, USA. [31]Seattle Children's Hospital Research Institute, 1900 9th Avenue, Seattle, Washington 98101, USA. [32]Genome Biology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Building 400-404, Walnut Creek, California 94598, USA. [33]Department of Plant Biology, University of Georgia, Athens, Georgia 30602-7271, USA. [34]Department of Molecular and Cell Biology, University of California, 539 Life Sciences Addition, Berkeley, California 94720-

3200, USA. [35]School of Computing Science, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. [36]Centre for Integrative Systems Biology of Ageing and Nutrition (CISBAN), Henry Wellcome Laboratory for Biogerontology Research, Newcastle University, Newcastle General Hospital, Newcastle upon Tyne NE4 6BE, UK. [37]Biological Data Management and Technology Center, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA. [38]Department of Ecology and Evolutionary Biology, University of California, 455 Steinhaus Hall, Irvine, California 92697, USA. [39]Molecular Infectious Diseases Group, Weatherall Institute of Molecular Medicine and University of Oxford Department of Paediatrics, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK. [40]Department of Biology, Howard University, 415 College Street, NW, Washington, DC 20059, USA. [41]LTER Network Office, Department of Biology, University of New Mexico, Albuquerque, New Mexico 87171, USA. [42]SIMBIOS Centre, University of Abertay Dundee, Dundee, DD1 1HG, UK. [43]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Mishima, Shizuoka 411-8540, Japan. [44]Center for Biological Sequence Analysis, The Technical University of Denmark, Lyngby, DK-2800 Kgs. Lyngby, Denmark. [45]Department of Molecular Biology, University of Wyoming, Laramie, Wyoming 82071, USA. [46]Center for Bioinformatics and Department of Genetics, University of Pennsylvania School of Medicine, 14th Floor Blockley Hall, 423 Guardian Drive, Philadelphia, Pennsylvania 19104, USA. Correspondence should be addressed to D.F. (dfield@ceh.ac.uk).

VI. A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML)

Authors: Renzo Kottmann, Tanya Gray, Sean Murphy, Leonid Kagan, Saul Kravitz, Thierry Lombardot, Dawn Field, and Frank Oliver Glöckner. 2008.

Published in *Omics*. 2008; 12(2): 101-8.

OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008

DOI: 10.1089/omi.2008.0A10

A Standard MIGS/MIMS Compliant XML Schema: Toward the Development of the Genomic Contextual Data Markup Language (GCDML)

Renzo Kottmann,[1] Tanya Gray,[2] Sean Murphy,[3] Leonid Kagan,[3] Saul Kravitz,[3] Thierry Lombardot,[1] Dawn Field,[2] Frank Oliver Glöckner,[1] and the Genomic Standards Consortium

Abstract

The Genomic Contextual Data Markup Language (GCDML) is a core project of the Genomic Standards Consortium (GSC) that implements the "Minimum Information about a Genome Sequence" (MIGS) specification and its extension, the "Minimum Information about a Metagenome Sequence" (MIMS). GCDML is an XML Schema for generating MIGS/MIMS compliant reports for data entry, exchange, and storage. When mature, this sample-centric, strongly-typed schema will provide a diverse set of descriptors for describing the exact origin and processing of a biological sample, from sampling to sequencing, and subsequent analysis. Here we describe the need for such a project, outline design principles required to support the project, and make an open call for participation in defining the future content of GCDML. GCDML is freely available, and can be downloaded, along with documentation, from the GSC Web site (http://gensc.org).

Introduction

It is well known that the entire collection of genomic and metagenomic DNA sequences is a complex and valuable resource (Field and Hughes, 2005). Hundreds of archaeal, bacterial, and eukaryotic genomes have been sequenced since the first microbial genome, *H. influenzae* (Fleischmann et al., 1995), was published a decade ago. Additionally, DNA sequences of thousands of organelles, plasmids, and viruses are available. In recent years, metagenomic sequencing has also become more prominent, with the largest single input of new sequences provided by the Global Ocean Survey (GOS) (Rusch et al., 2007). Taking into account the immense genomic diversity found in the natural world (Binnewies et al., 2006), it is anticipated that large-scale metagenomic sequencing projects, such as GOS, flag just the beginning of a new era in molecular microbiology.

In particular, in light of the rapidly growing number of environmental and microbiome sequencing projects, it is increasingly clear that the biological interpretation of such data, especially in a comparative context, is dependent on the quantity and quality of associated information (Field et al., 2008a; Raes et al., 2007). It is the aim of the Genomic Standards Consortium (GSC) to support the capture of a richer set of data. The first step of this international community has been to define the "Minimum Information about a Genome Sequence" (MIGS) and "Minimum Information about a Metagenome Sequence" (MIMS) specifications. Use of MIGS/MIMS will provide a mechanism for capturing a consensus-driven minimum set of metadata describing aspects of genomes and metagenomes, such as geographic location and habitat type from which the sample was taken, as well as the details of the sequencing method used. The support of maximum reporting of such projects, though, will require a much richer set of descriptors (Raes et al., 2007). Such descriptors must cover both the origin and processing of a sample, from the time of sampling up to sequencing, and the subsequent analysis. This suite of metadata is collectively referred to here as contextual data.

It is the aim of the GSC to provide support for the capture of richer contextual data describing genomes and metagenomes by developing the Genomic Contextual Data Markup Language (GCDML). This project is the natural extension of original efforts to implement the MIGS/MIMS check-list as an XML Schema (MIGS.xsd) (Field et al., 2008a). The scope of this restricted schema evolved into the scope of GCDML to specifically support "maximal" reporting of contextual data and the desire of groups in the GSC to include local descriptors in the original MIGS/MIMS schema.

There are two key aspects to the development of GCDML. First are the technical aspects of "how to build it." Second is the issue of "scope," or "what to put in" (the exact descriptors to be included). Here, the focus is primarily on the former and outline the design principles of GCDML as a technical shell for future content development in the coming years. The current core scope of GCDML has been defined

[1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, 28359 Bremen, Germany.
[2]NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford, Oxfordshire, United Kingdom.
[3]J. Craig Venter Institute, Rockville Maryland.

by the MIGS/MIMS checklist, and covers the minimum description of "nucleic acid sequence source," "environment," and "sequencing methodology" (Field et al., 2008a). Beyond that, it will take future GSC workshops, telecoms, and significant participant contribution to develop the full content of GCDML based on the needs and interests of the community.

The Scope of GCDML and Design Considerations

GCDML aims to take full advantage of the benefits of an XML representation of genomic contextual data. XML provides a machine-readable representation of metadata that facilitates the capture, exchange, and comparison of large amounts of data. XML is widely used to describe data capture and exchange format. Numerous XML definitions exist in bioinformatics (for an overview, see Seibel et al., 2006). XML definitions such as BSML (Cerami, 2005) and INSDSeq (http://www.insdc.org/page.php?page=documents last verified on 15.02.2008) have applications in sequence annotation; PSAML (Su-Hyun et al., 2002), ProML (Hanisch et al., 2002) and PSI MI (Hermjakob et al., 2004) in protein modeling and interaction; MAGE-ML (Spellman et al., 2004) in gene expression. The Functional Genomics (FuGE) project aims to develop a single generic data model that will underpin a variety of XML-based formats by providing a single common framework (Jones et al., 2008). XML is also widely used in many other scientific fields, and for example, the Ecological Modeling Language standard is described in this issue (San Gil et al., 2008).

In overview, MIGS/MIMS will be central to GCDML, and GCDML will provide the GSC's official implementation of the checklist. This requires GCDML to specifically support the validation of different subsets of descriptors because MIGS/MIMS is applied differently across taxa (e.g., eukaryotes vs. bacteria vs. viruses or metagenomes) (Field et al., 2008a). Beyond the minimum descriptors of MIGS/MIMS, GCDML will be open and extensible to evolve with the needs of the community. In mature versions of GCDML, it is envisioned that it will be possible to describe the exact origin and processing of a biological sample from the time of sampling up to sequencing, and subsequent analysis. This should support a range of desired applications, such as tracking the geographic origin and habitat of a sample or a set of organisms, for example, comparative ecological genomic studies, to capture the pathogenicity of a sequenced organism, or to describe the host–microbiome relationship in human microbiome studies. GCDML should also be viewed as a single, community developed specification for exchanging data between databases and providing integrated information from the resources to the wider community.

At the technical level, GCDML must be designed to enable support for the above requirements, facilitate a broad adoption, and support the future inclusion of a far wider range of descriptors. GCDML is therefore being built to have a strongly typed and clear structure that embodies the following design principles. GCDML must be:

- "application agnostic" (to facilitate the mapping of data from diverse resources and applications, and the use by applications with their own needs, e.g., Web services)
- compliant with (MIGS/MIMS), while allowing for richness of expression
- support the integration of terms lists, including those from ontologies
- allow the recording of legacy data, even when fields are missing
- open and extensible to allow evolution of the MIGS/MIMS specification and all associated optional descriptors of genomes and metagenomes, as well as the evolution of new databases and sources of related metadata.
- open to link, map, or incorporate other standards as required, and in particular, provide integration with of Geography Markup Language (GML), which became ISO standard 19136
- support versioning of GCDML and any reports generated from it.

These complex requirements first necessitated a shift in the design of the prototype MIGS.xsd schema from a Russian doll model to an open flat design (van der Vlist, 2002), which makes the schema highly modular (no nesting of elements). The hierarchical structure of the different report types is subsequently created by appropriately nesting references to the existing elements (see below). This provides significant practical benefits. First, it facilitates implementation of one kind of genome report for all types of MIGS/MIMS (eukaryote, bacteria and archaea, plasmid, virus, organelle, or metagenome) project types (Field et al., 2008a). Now, any report can, using the relevant subsets of MIGS/MIMS descriptors, be validated by any XML Schema-capable XML parser. Second, it allows definition of additional descriptors, which do not interfere with MIGS/MIMS genome reports. Third, the flat design is the basis for an open, extensible design (see below), which allows both the necessary evolution of the MIGS/MIMS specification and the additional descriptors (Field et al., 2008a).

Supporting MIGS/MIMS compliance: report structure

Although reports can drastically differ in their details, as they do for MIGS/MIMS compliant reports, in GCDML they all follow the same semantic structure. The main XML elements are simplified models of samples that are produced during genomic studies. The order of the XML elements follows a simplified and generalized protocol for sample collection (<originalSample>), in the case of a single organism, the isolation (<isolate>), DNA extraction (<dnaExtract>), DNA library construction (<DNALibrary>), and sequencing details, including assembly (<sequencing>). The descriptors of the MIGS/MIMS checklist are associated with the main XML elements via nesting. An example, as depicted in Figure 1, is the nesting of sampling time (<samplingTime />), sample location (<_SampleLocation />), and habitat (<_Habitat />) within <_originalSample />.

MIGS/MIMS-compliant reports share the same root element

The nasReports (Nucleic Acid Sequence reports) element is the root element of all five MIGS/MIMS compliant report types for Eukarya, Bacteria, Archaea, Plasmids, Viruses, Organelles, and Metagenomes (Fig. 1). The root element acts as

```
<nasReports>
  <_Report>
    <gcatID />
    <studyData><extension /></studyData>
    <originalSample>
      <samplingTime />
      <_SampleLocation><extension /><_SampleLocation>
      <_Habitat ><extension /></_Habitat >
      <extension/>
    </originalSample>
    <isolate><extension /></isolate>
    <dnaExtract><extension /></dnaExtract>
    <dnaLibrary><extension /></dnaLibrary>
    <sequencing><extension /></sequencing>
    <extension />
  </_Report>
</nasReports>
```

FIG. 1. General structure of nucleic acid sequence (nas) reports including extension points (<extension/>) within various elements.

a container for any number of NAS Reports per XML document. GCDML does not put any constraint on this container. Thus, NAS reports can be grouped in any way, for example, from groups of different reports of the same genome (but from different sources), to groups of genome reports of the same taxonomic group.

Keeping "free text" to a minimum and the use of categorical terms lists for report values

Validation of categorical terms with standard XML parsers is achieved through the enumeration of valid terms. XML schema allows the construction of simple types with a restricted sets of terms, which can come from ontologies, taxonomies, gazetteers, and other sources of controlled vocabularies. This general approach allows mixing of terms from different sources. GCDML makes wide use of categorical terms for many MIGS/MIMS descriptors. An XML schema example is given in Figure 2a, which restricts an XML type to the terms "soil" and "water," as depicted in Figure 2d. This technique allows syntactical restriction of the set of useful terms and validation with standard XML parsers. Therefore, the use of free-text fields is decreased to an absolute minimum.

Supporting the use of ontologies for categorical terms

The GSC is strongly committed to the use and development of ontologies. For example, the GSC is a member community in the Ontology for Biomedical Investigations (Whetzel et al., 2006), a founding member of the is Environment Ontology project (http://environmentontology.org), and is driving the development of a minimum controlled vocabulary of habitat terms (Hirschman et al., 2008). The GSC has agreed that, just as for MIGS/MIMS (Field et al., 2008a), there is a need to provide semantic transparency for GCDML through the integration of ontologies. An outstanding issue is how to do it best. The following consensus emerged from discussions within the GSC: finding a solution that does not add complexity to NAS reports, does not depend on the availability of ontologies, and does not force users of GCDML, and/or NAS reports, to become acquainted with ontologies.

GCDML currently uses SAWSDL, which stands for "Semantic Annotation for WSDL and XML Schema" (Kopecký et al., 2007). The World Wide Web Consortium recommendation (W3C; August 2007) allows annotation of XML Schema with references to ontological concepts (independent of the type and format of the ontology). SAWSDL allows separation of the syntactic modeling of data and semantic modeling of a knowledge domain by first focusing on the use of XML Schema to model data. Next, SAWSDL is used to state within the GCDML schema which XML schema construct has a meaningful relationship to which ontological entity. Thus, enumeration of categorical terms can be used to ensure the syntactic consistency of NAS reports, while allowing semantic applications to utilize NAS reports with ontological concepts. Figure 2c shows the XML Schema from Figure 2a with SAWSDL annotation, which still validates the same XML documents as the non-annotated XML schema. SAWSDL can be introduced at any time, because it would not affect NAS reports in any way.

Integration of legacy data with "missing fields"

The question of how to integrate noncompliant legacy data, without relaxing rules and constraints on future data incorporation, is an important issue in the application of any new standard. A case study is the issue of reporting sample location and sampling time, which are mandatory according to the MIGS/MIMS check-list (Field et al., 2008a). Although the community agreed it is necessary to require sample location and time in the future, it was realized that these data are often not available for legacy data.

To overcome this, XML Schema allow for the creation of unions of simple types. GCDML introduces a special simple type for the enumeration of categorical terms that indicate where and why data is missing. A simplified example is given in Figure 2b: first, a simple type, "null," is created; second, a new simple type, "geoFeatureUnion," is created as the union of the "null" simple type and the "geographicFeature" simple type, as given in Figure 2a. GCDML uses these union types to explicitly state reasons for missing data throughout the schema.

Openness and extensibility of GCDML

The extensibility of GCDML is defined by the ability of other applications to use elements from GCDML as a basis for its own element definitions (van der Vlist, 2002). This is facilitated by the flat design of GCDML, use of substitution groups, and almost exclusive derivation by extension (van der Vlist, 2002). Openness of an XML schema is achieved if the schema allows addition of any content at well-defined extension points in XML documents. GCDML adds extension points to <nasReports>, <_Report>, <studyData>, <sequencing>, <habitat>, and each element that models a sample (see Fig. 1). This allows applications generating MIGS/MIMS compliant <nasReports> to add additional self-defined XML within its extension points, which adhere to GCDML.

a) Enumeration of categorical terms

```
<simpleType name="geographicFeature">
  <restriction base="string">
    <enumeration
        value="soil" />
    <enumeration
        value="water" />
  </restriction>
</simpleType>

<element name="GeoFeature1"
    type=" geographicFeature" />
```

b) Explication of unknown values

```
<simpleType name="null">
  <restriction base="string">
     <enumeration
         value="unknown" />
     <enumeration
         value="inapplicable" />
  </restriction>
</simpleType>

<simpleType name="geoFeatureUnion">
  <union memberTypes="null
            geographicFeature" />
</simpleType>

<element name="GeoFeature2"
     type=" geoFeatureUnion" />
```

c) Annotation with SAWSDL

```
<simpleType name="geographicFeature">
  <restriction base="string">
    <enumeration
        value="soil"
        modelReference=
"http://purl.org/obo/owl/ENVO#ENVO_00001998" />
    <enumeration
        value="water"
        modelReference=
"http://purl.org/obo/owl/ENVO#ENVO_00002006" />
  </restriction>
</simpleType>
```

d) XML documents defined in a)

```
<GeoFeature1>soil</GeoFeature1>
<GeoFeature1>water</GeoFeature1>
```

e) XML documents defined in b)

```
<GeoFeature2>soil</GeoFeature2>
<GeoFeature2>water</GeoFeature2>
<GeoFeature2>unknown</GeoFeature2>
<GeoFeature2>inapplicable </GeoFeature2>
```

FIG. 2. XML examples demonstrating different implementation approaches of GCDML: (**a**) a simplified XML schema using enumerations to restrict values to the categorical terms "hot spring" and "hydrothermal vent" only; (**b**) a simplified XML Schema for handling missing data; (**c**) the same simpletype as in (**a**) annotated following the SAWSDL standard; (**d**) simplified XML documents showing all possible uses of <GeoFeature1> as defined in (**a**); and (**e**) simplified XML documents showing all possible uses of <GeoFeature2> as defined in (**b**).

Versioning GCDML and NAS reports

Versioning is the process of assigning unique version numbers to either unique states of GCDML or unique states of NAS reports, and other applications of GCDML. However, versioning GCDML is separate from versioning NAS reports in the sense that a version of GCDML indicates a certain state of the grammatical structure and constraints that apply to NAS reports. Whereas versioning of NAS reports indicate the state of the information content of the report.

GCDML will use the common *major.minor.release* versioning scheme, with version "1.5.0" serving as the first "public" release version. The criterion for an increase of the *major* number is a feature extension to GCDML. A *minor* version update will be triggered by a set of changes to GCDML that require NAS reports to be transformed in order to concur with the new version. Finally, the *release* number is set to increase in the case of changes to GCDML that have no effect on any existing NAS report.

Early Adopters of GCDML

Several groups presented on their intentions to adopt GCDML at the fifth GSC workshop (Field et al., 2008c). The first true example of adoption of GCDML within a database is that of the GSC's Genome Catalog (GCat). GCat is an online database designed to collect MIGS/MIMS compliant reports. It is run with the generic GenCat software, which creates an online database system with data input, edit, browse, and search functions from XML Schemas to allow capture of XML schema-compliant reports. Adoption has involved replacement of the original MIGS.xsd with the gcdml.xsd and the addition of support for new features found in GCDML. GCat will help the GSC to demonstrate and clarify the details of GCDML to the wider community, as it renders the GCDML as an input form. The GCDML-driven input form identifies optional elements, repeatable elements, substitution groups, and choices, as well as lists the values of enumerated elements. In addition, element names and associated documentation contained in the XML schema are displayed in the input form. GCat further supports the ongoing development of GCDML by, for example, its support of term capture. This will allow newly submitted terms in the input form to be included as enumerated elements in GCDML. The terms can later be reviewed by a third party, for potential inclusion in the next revision of GCDML.

As described in this special issue, the use of GCDML is now also being explored by the Long-Term Ecological Research (LTER) Network (San Gil et al., 2008). The LTER is a collaborative effort funded by the National Science Foundation to investigate ecological processes over long temporal and broad spatial scales. Increasingly, the microbial observatories within the LTER are generating genomic and metagenomic data. The LTER has produced the Ecological Metadata Language (EML) standard, an XML Schema that provides a metadata specification for describing data relevant to the ecological discipline (Fegraus et al., 2005).

A Roadmap for GCDML and an Open Call for Participation

GCDML was initially proposed at the fourth GSC Workshop in June 2007 (Field et al., 2008b). Subsequently, a GCDML prototype was developed and further improvements made based on community feedback, including that of experts on the biology of particular types of genomes. At the fifth GSC Workshop, GCDML was officially accepted as the GSC implementation of MIGS/MIMS and as a candidate standard mechanism for data exchange across a range of databases (Field et al., 2008c). As the next top-priority core GSC project, following the publication of the MIGS specification (Field et al., 2008c), our top aim is to develop a stable release of GCDML and additional working examples of the use of GCDML within the community.

The construction of GCDML will largely progress in three overlapping phases:

1. Setup of GCDML with a primary focus on supporting the creation of MIGS/MIMS-compliant XML reports
2. Integration of ontologies
3. Enlargement of the set of descriptors based on the needs of a variety of databases and additional specifications (e.g., MINIMESS)

Phase 1 is largely complete, and MIGS/MIMS version 2.0 is implemented in the current schema. Phase 2 is in the early stages, and will depend heavily on the maturation of a variety of key controlled vocabularies, like Habitat-Lite (Hirschman et al., 2008), and ontologies (e.g., the Environment Ontology (http://environmentontology.org) upon which Habitat-Lite is based). Phase 3 will start in the summer of 2008.

A key purpose of the description of GCDML presented here is to serve as an open call for participation in the GCDML project. Just as for the GSC as a whole (Field et al., 2008a) our activities are open to new contributors and end users at any time; only through wide participation of the community will GCDML evolve to be a valuable consensus-driven tool of use beyond the GSC's Genome Catalog. In particular, this work should proceed through the development of a series of case studies. As GCDML matures, we aim to shift to a more formal mechanism of vetting the inclusions of terms and changes to the core schema. We plan to produce a fully MIGS/MIMS compliant version of GCDML with support for ontologies by the end of 2008 for presentation at the combined GSC/"Metagenomics 2008" meeting. All versions of the GCDML schema and additional documentation are available at http://sourceforge.net/projects/gensc.

Discussion

GCDML meets the two key goals of the GSC, providing both richer descriptions of genomes and metagenomes and facilitating open and transparent data exchange (Field et al., 2008a). The purpose of GCDML is to provide a mechanism for capturing MIGS/MIMS-compliant reports and a larger XML vocabulary for transparent contextual data exchange between specialized local databases, such as the GSC's Genome Catalog, Genome Reviews (Sterk et al., 2006), GOLD (Liolios et al., 2008), IMG (Markowitz et al., 2008), Genomes Mapserver (Lombardot et al., 2006), and CAMERA (Seshadri et al., 2007). The technical design principles outlined here will enable this in the future.

The sample-centric design of the NAS Reports is in accordance with the harmonization efforts of the different standardization initiatives of the "omics" community (Morrison et al., 2006), especially the "Investigation, Study, Assay" concept of the Reporting Structures for Biological Investigations (RSBI) working group (Sansone et al., 2008). Moreover, the sample-centric structure of NAS reports (Fig. 1) has proven to be quite stable through a round of several iterations of the schema. The numerous improvements of GCDML based on community feedback from experts on the biology of particular types of genomes did not change the general structure.

The sample-centric, open, and extensible design of GCDML serves well as the basis for applications beyond MIGS/MIMS-compliant reports as well. For example, at the last GSC meeting, members of the Alpine Microbial Observatory proposed the extension of GCDML to report contextual data for single rRNA sequences (Field et al., 2008c). This would undoubtedly benefit the "molecular diversity" community, which is faced with more than 500,000 rRNA sequences in the public nucleotide databases (Cole et al., 2007; DeSantis, 2006; Pruesse et al., 2007).

Further community acceptance of GCDML can be expected through the integration of existing industrial strength standards, such as Geography Markup Language (Lake et al., 2004) and SAWSDL (Kopecký et al., 2007). Advanced programming interfaces exist for SAWSDL to facilitate ontological use of GCDML and many geographical tools already support GML and derived standards (http://www.opengeospatial.org/resource/products, last verified on 15.02.2008). Besides public visibility, one key to broad acceptance of GCDML is comprehensive user and developer documentation, including cookbook style best practice guides.

Summary

GCDML serves as the official XML schema implementation of the MIGS/MIMS specification, as defined by the Genomic Standards Consortium (GSC) (Field et al., 2008a). The design of GCDML is guided by several principles. First, it is strongly typed, avoids free-text fields through the use of enumerations, and supports the use of terms from ontologies. This is expected to have a significant impact on data integrity. Second, GCDML complies with the minimum requirements of MIGS/MIMS, but additionally allows a much richer contextual description of genomic studies. Third, it allows the incorporation of nonstandard legacy data without weakening the strong typing. Fourth, the flat design allows other applications to use selected parts of GCDML for their own needs, and is therefore well suited, for example, for

Web-service applications. Finally, the open and extensible design of GCDML allows feature extensions with minimal effort to update existing reports. Thus, making GCDML prepared for extensions, such as the inclusion of descriptors proposed by MINIMESS (Raes et al., 2007) and additional descriptors for more comprehensive description of the annotation processes of genomic and metagenomic sequences.

Acknowledgments

We would like to thank the experts on the biology of particular types of genomes and metagenomes for their valuable comments during the development of GCDML, namely, Jack Gilbert, Keith James, Gareth Wilson, Sarah Turner, Christiane Hertz-Fowler, Peter Dawyndt, Phillip Goldstein, and Melissa Duhaime. We thank the EU Sixth Framework Programme (FP6-NEST) for providing financial support (MetaFunctions project, contract no. 511784). Tanya Gray is support by NERC Grants NE/E007325/1 and NE/3521773/1 to D.F.

Author Disclosure Statement

The authors declare that no competing finacial interests exists.

References

Binnewies, T.T., Motro, Y., Hallin, P.F., Lund, O., Dunn, D., La, T., et al. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries.. Funct Integr Genomics 6, 165–185.

Cerami, E. (2005). *XML for Bioinformatics* (Springer Science+ Buisness Media, Inc., New York).

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., et al. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res 35, D169–D172.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72, 5069–5072.

Fegraus, E.H., Andelman, S., Jones, M.B., and Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bull Ecolo Soc Am 86, 158–168.

Field, D., and Hughes, J. (2005). Cataloguing our current genome collection. Microbiology 151, 1016–1019.

Field, D., Garrity, G.M., Gray, T., Morrison, N., Selengut, J.D., Sterk, P., et al. (2008a). The "Minimum Information about a Genome Sequence"(MIGS) specification. Nat Biotechnol 26, 541–547.

Field, D., Glöckner, F.O., Garrity, G.M., Gray, T., Sterk, P., Cochrane, G., et al. (2008b). Meeting report: the 4th Genomic Standards Consortium (GSC) workshop. OMICS (this issue).

Field, D., Garrity, G.M., Sansone, S.-A., Sterk, P., Gray, T., and Glöckner, F.O. (2008c) The fifth Genomic Standards Consortium Workshop meeting report. OMICS (this issue).

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Hanisch, D., Zimmer, R., and Lengauer, T. (2002). ProML—the protein markup language for specification of protein sequences, structures and families. In Silico Biol 2, 313–324.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., et al. (2004). The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat Biotechnol 22, 177–183.

Hirschman, L., Clark, C., Bretonnel, C., Mardis, S., Luciano, J., Cole, J., et al. (2008). Habitat-Lite: a GSC case study based on free text terms for environmental metadata. OMICS (this issue).

Jones, A.R., Miller M., Aebersold, R., Apweiler, R., Ball, C.A., Brazma, A., et al. (2008). The functional genomics experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol 25, 1127–1133.

Kopecký, J., Vitvar, T., Bournez, C., and Farrel, J. (2007). SAWSDL: semantic annoations for WSDL and XML schema. IEEE Internet Comput 11, 60–67.

Lake, R., Burggraf, D., and Trninic, M. (2004). *Geography Mark-Up Language (GML): foundation for the Geo-Web* (John Wiley & Sons Ltd., West Sussex, London).

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, D475–D479.

Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C. et al. (2006). Megx.net—database resources for marine ecological genomics. Nucleic Acids Res 34, D390–D393.

Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.A., et al. (2008). The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res 36, D528–D533.

Morrison, N., Cochrane, G., Faruque, N., Tatusova, T., Tateno, Y., Hancock, D., et al. (2006). Concept of sample in OMICS technology. OMICS 10, 127–137

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188–7196.

Raes, J., Foerstner, K.U., and Bork, P. (2007). Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10, 490–498.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5, e77.

San Gil, I., Sheldon, W., Schmidt, T., Servilla, M., Aguilar, R., Greis, C., et al. (2008). Defining linkages between the GSC and NSF's LTER program: how the ecological metadata language relates to GCDML and other outcomes. OMICS (this issue).

Sansone, S.-A., Rocca-Serra, P., Brandizi, M., Brazma, A., Field, D., and Fostel, J. (2008). The first RSBI (ISA-TAB) workshop: "Can a simple format work for complex studies?" OMICS (this issue).

Seibel, P.N., Krüger, J., Hartmeier, S., Schwarzer, K., Löwenthal, K., Mersch, H., et al. (2006). XML schemas for common bioinformatic data types and their application in workflow systems. BMC Bioinformatics 7, 490.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: a community resource for metagenomics.. PLoS Biol 5, e75.

Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., et al. Design and implementation

of microarray gene expression markup language (MAGE-ML). Genome Biol. 3, 9.
Sterk, P., Kersey, P.J., and Apweiler, R. (2006). Genome reviews: standardizing content and representation of information about complete genomes. OMICS 10, 114–118.
Su-Hyun, L., Jin-Hong, K., Geon-Tae, A., and Myung-Joon, L. (2002). An XML representation of protein data for efficient structure comparison. *Proc. of Second International Conference on Computer and Information Science*, No. 1, p. 313.
van der Vlist, E. (2002). *XML Schema*. (O'Reilly & Associates, Inc., Sebastopol, CA).
Whetzel, P.L. Brinkman, R.R., Causton, H.C., Fan, L., Field, D., Fostel, J., et al. (2006). Development of FuGO: an ontology for functional genomics investigations. OMICS 10, 199–204.

Address reprint requests to:
Renzo Kottmann
Microbial Genomics Group
Max Planck Institute for Marine Microbiology and
Jacobs University Bremen
Celsiusstr. 1
28359 Bremen, Germany

E-mail: rkottman@mpi-bremen.de

VII. Defining Linkages between the GSC and NSF's LTER Program: How the Ecological Metadata Language (EML) Relates to GCDML and Other Outcomes

Authors: Inigo San Gil, Wade Sheldon, Tom Schmidt, Mark Servilla, Raul Aguilar, Corinna Gries, Tanya Gray, Dawn Field, James Cole, Jerry Yun Pan, Giri Palanisamy, Donald Henshaw, Margaret O'Brien, Linda Kinkel, Katherine McMahon, Renzo Kottmann et al.

Published in *Omics.* 2008; 12(2):151-6.

Contribution: Evaluation and discussion of the relation between EML and GCDML.

OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008

DOI: 10.1089/omi.2008.0015

Defining Linkages between the GSC and NSF's LTER Program: How the Ecological Metadata Language (EML) Relates to GCDML and Other Outcomes

Inigo San Gil,[1] Wade Sheldon,[2] Tom Schmidt,[3] Mark Servilla,[1] Raul Aguilar,[4] Corinna Gries,[4] Tanya Gray,[5] Dawn Field,[5] James Cole,[3] Jerry Yun Pan,[6] Giri Palanisamy,[6] Donald Henshaw,[7] Margaret O'Brien,[8] Linda Kinkel,[9] Katherine McMahon,[10] Renzo Kottmann,[11] Linda Amaral-Zettler,[12] John Hobbie,[13] Philip Goldstein,[14] Robert P. Guralnick,[14] James Brunt,[1] and William K. Michener[1]

Abstract

The Genomic Standards Consortium (GSC) invited a representative of the Long-Term Ecological Research (LTER) to its fifth workshop to present the Ecological Metadata Language (EML) metadata standard and its relationship to the Minimum Information about a Genome/Metagenome Sequence (MIGS/MIMS) and its implementation, the Genomic Contextual Data Markup Language (GCDML). The LTER is one of the top National Science Foundation (NSF) programs in biology since 1980, representing diverse ecosystems and creating long-term, interdisciplinary research, synthesis of information, and theory. The adoption of EML as the LTER network standard has been key to building network synthesis architectures based on high-quality standardized metadata. EML is the NSF-recognized metadata standard for LTER, and EML is a criteria used to review the LTER program progress. At the workshop, a potential crosswalk between the GCDML and EML was explored. Also, collaboration between the LTER and GSC developers was proposed to join efforts toward a common metadata cataloging designer's tool. The community adoption success of a metadata standard depends, among other factors, on the tools and trainings developed to use the standard. LTER's experience in embracing EML may help GSC to achieve similar success. A possible collaboration between LTER and GSC to provide training opportunities for GCDML and the associated tools is being explored. Finally, LTER is investigating EML enhancements to better accommodate genomics data, possibly integrating the GCDML schema into EML. All these action items have been accepted by the LTER contingent, and further collaboration between the GSC and LTER is expected.

Background

The LTER and the Genomics Standards Consortium (GSC) initiated joint discussions in November 2007 during a Long-Term Ecological Research Network (LTER) meeting held at Michigan State University (MSU). James Cole, head of the Ribosomal Database Project, gave an overview of the difficulties associated with annotating 16S sequences,

[1]Department of Biology, LTER Network Office, University of New Mexico, Albuquerque, New Mexico.
[2]Department of Marine Sciences, University of Georgia, Athens, Georgia.
[3]Department of Microbiology & Molecular Genetics, Michigan State University, East Lansing, Michigan.
[4]Global Institute of Sustainability, Arizona State University, Tempe, Arizona.
[5]NERC Centre for Ecology and Hydrology, Oxford, OX1 3SR, United Kingdom.
[6]Oak Ridge Natl. Lab, Oak Ridge, Tennessee.
[7]USDA Forest Service, Pacific NW Research Station, Corvallis, Oregon.
[8]Marine Science Institute, University of California at Santa Barbara, Santa Barbara, California.
[9]Department of Plant Pathology, University of Minnesota, Saint Paul, Minnesota.
[10]Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, Wisconsin.
[11]Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, Bremen, Germany.
[12]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts.
[13]The Ecosystems Center, Marine Biological Laboratory, Woods Hole, Massachusetts.
[14]Department of Ecology and Evolutionary Biology and University of Colorado Museum of Natural History, University of Colorado, Boulder, Colorado.

genomes, and metagenomes with contextual information and outlined current GSC activities in the area. Significant discussion followed, and given that the GSC already had an interest in developing linkages to researchers generating genomes and metagenomes in the LTER, the LTER was invited to send a representation to the fifth GSC workshop. In particular, the GSC was keen to understand how its Genomic Contextual Data Markup Language (Kottman et al., 2008) (GCDML) might relate to the LTER's Ecological Metadata Language (EML) (Jones et al., 2001). The GSC also strives to learn from the experience of other communities and was therefore keenly interested in how the best practices of the LTERs were applied in building EML and how it achieved NSF adoption of EML. Finally, the GSC was interested, through these linkages, to increase participation of LTER members in GSC goals. This report details these discussions, and elaborates on subsequent interactions that are helping to define how and why these two communities should work together further in the future.

LTER input into the fifth GSC workshop

At the workshop, Inigo San Gil presented a talk that covered essential background to the LTER Network, EML, and shared experience on developing this XML standard, gaining adoption, and general issues of data integration.

The LTER

The LTER was formed in 1980 with support from the NSF. The LTER today has over 2000 researchers associated with 26 sites representing diverse ecosystems and research emphases in continental North America, islands in the Caribbean, the Pacific, and Antarctica—including deserts, estuaries, lakes, oceans, coral reefs, prairies, forests, alpine, and Arctic tundra, urban areas, and production agriculture. The LTER mission is to provide the scientific community, policy makers, and society with the knowledge and predictive understanding necessary to conserve, protect, and manage the nation's ecosystems, their biodiversity, and the services they provide. When the first six LTER sites in the LTER Network were funded in 1980, the idea of studying specific ecosystems at temporal and spatial scales was revolutionary. Currently, all 26 LTER sites participate in this endeavor, and there are over 30 other countries that formed LTER networks with presence in all continents. Many other countries are in the process of forming LTER networks, most of them following similar procedures and frequently adopting the same metadata standards as in the US LTER. To learn more about the status of the non-US LTERs, one can google the Internet, that is, international LTER networks.

One of the challenges that the LTER Network has faced since its inceptions has been data synthesis. LTER started generating and offering to the public metadata and data very early on (Michener et al., 1997); however, different LTER sites adopted different protocols to divulge their metadata. Researchers focusing on cross-site comparisons faced formidable challenges to pursue synthesis projects. To address some of these challenges, LTER adopted a network-wide metadata standard in 2001 and gradually has standardized nearly all existing metadata records (San Gil and Baker, 2007). The LTER Network metadata standard, EML, is recognized by NSF (LTER Information Managers Executive Committee, 2005), and EML rules compliance is used by the NSF review teams to evaluate each LTER site's performance. Particularly, the official NSF LTER site review criteria document states that: "(1) Metadata shall be of sufficient quality and completeness to ensure long-term (>20 years) usability of data. (2) Metadata shall be EML-compliant at level 2 (a document shall be discoverable; metadata content is well beyond the minimums dictated by the EML schema rules). Metadata should be EML-compliant at level 5 (level 5 is also known as the integration level, when the metadata enables machine readability for associated data). LTER Site EML shall comply with LTER best practices." Note that positive site reviews are critical for continuing funding of the on-going long-term programs at each LTER site.

The EML

EML (Jones et al., 2001) is a comprehensive standard that has been adopted by a sector of the larger international ecological research community. EML has had two major revisions, and the development committee is currently working on a new release. Even though the EML standard enables data integration at the machine level (with little or no human intervention), it is thought to lack adequate information holders for more specialized genomic datasets.

EML is implemented as a series of XML schema document types that can by used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological dataset. EML has four general descriptors at the top of the hierarchy. One can choose to describe a dataset, a protocol, a citation, or software. The EML-dataset module contains general information that describes dataset resources. It is intended to provide overview information about the dataset: broad information such as the title, abstract, keywords, contact information, maintenance history, and distribution of the data themselves. The EML-dataset module also imports many other modules that are used to describe the dataset in fine detail. Specifically, it uses the EML-methods module to describe methodology used in collecting or processing the dataset, the EML-project module to describe the overarching research context and experimental design, the EML-entity module to provide detailed information about the logical structure of the dataset. A dataset often is composed of a series of data entities (tables, shape files, photos, and the like) that are linked together by particular integrity constraints. The EML-literature module contains information that describes literature resources. It is intended to provide overview information about the literature citation, including title, abstract, keywords, and contacts. Citation types include: article, book, chapter, edited book, manuscript, report, thesis, conference proceedings, personal communication, map, generic, audio visual, and presentation. The "generic" citation type would be used when one of the other types will not work. The EML-software module contains general information that describes software resources, while the EML-protocol module is used to define abstract and prescriptive procedures for the dataset.

Other EML modules that are not at the top of the EML hierarchy, but can be used in several contextual places in a dataset are the EML-coverage, EML-data table, EML-spatial raster or vector, EML-physical, and EML-attribute modules.

There are more EML modules, but for the sake of briefness, we shall focus only on these four. The EML-coverage module contains fields for describing the coverage of a resource in terms of time, space, and taxonomy. These coverages (temporal, spatial, and taxonomic) represent the extent of applicability of the resource in those domains. The geographic coverage is expressed via a set of bounding coordinates that define the North, South, East, and West points in a rectangular area, optionally including a bounding altitude, or using a G-Ring polygon definition, where an irregularly shaped area may be defined using a ordered list of latitude/longitude coordinates. The temporal coverage section allows for the definition of either a single date/time, or a range of dates/times expressed according to the ISO 8601 Date and Time Specification. The taxonomic coverage section allows for detailed description of the taxonomic extent of the dataset or resource.

The EML-dataTable module is used to describe each tabular set of information in a dataset. A series of comma or tab-separated text files may be considered a dataset, and each file would subsequently be considered a data table entity within the dataset. Data tables may be ASCII text files, relational database tables, spreadsheets, or other type of tabular data with a fixed logical structure. The EML-dataTable module allows for the description of each attribute (column/field/variable) within the data table through the use of the EML-attribute module. Likewise, there are fields used to describe the physical distribution of the data table, its overall coverage, the methodology used in creating the data, and other logical structure information such as its orientation, case sensitivity, etc. The EML-attribute module describes all attributes (variables) in a data entity: dataTable, spatialRaster, spatialVector, and other EML-entity modules. The description includes the name and definition of each attribute, its domain, definitions of coded values, and other pertinent information, including child modules to describe units, precision, quality controlled procedures, missing codes and the like. An attribute-list encompasses a list of attributes that go together in some logical way, while the attribute structure within the attribute list is used to define a single attribute.

The EML-physical module describes the external and internal physical characteristics of a data object as well as the information required for its distribution, that is, filename, delimiters, encoding methods, header/footer lines, authentication of a file, etc. Distribution information describes how to retrieve the data object. The retrieval information can be either online (e.g., a URL) or offline (e.g., a data object residing on an archival tape) or inline (attached to the metadata document). The EML-spatialRaster and the EML-spatialVector modules allow for the description of entities composed of data values that are usually georeferenced to a portion of the Earth's surface. Specific attributes of a spatial raster and vector can be documented here including the spatial organization of the raster cells, the cell data values, vectors, and the relationships among them.

Specific input into MIGS/MIMS and GCDML

Inigo San Gil contributed directly to discussion at the workshop on finalizing the "Minimum Information about a Genome/Metagenome Sequence" (MIGS/MIMS) specification (Field et al., 2008). He also provided considered suggestions about the nascent GCDML project on behalf of the EML project. Specifically, he expressed concerns about committing to a finite set of data parameters to describe a habitat. A workaround is to leave open the number of parameters, allowing the user to enter the parameter name and definition. He also suggested that GCDML currently lacks a place holder for additional metadata. That is, sometimes the researcher may want to document some details that cannot be captured elsewhere in the specification. A possible alternate solution is to include a place holder for additional information or additional metadata. Other suggestions for aiding the adoption of MIGS/MIMS and GCDML were drafting "best practices" documents, as well as proposing sponsored trainings, and finally, providing assistance in developing of customized GCDML compliant metadata ingestion tools through cooperation.

He also pointed out that here is a potentially synergistic overlap between LTER and GSC activities with respect to their choice of software development projects to enable the ingestion and export of data and metadata. The GSC and LTER communities have independently chosen the same technologies for the task to build metadata capture tools. These XML based tools are centred on Orbeon and XForms (Ajax, or the engines behind Web 2.0). The GSC's GenCat (Field et al., 2006) software takes an XML schema (e.g., originally the MIGS.xsd schema and now GCDML) and autogenerates an online data capture and repository system on the fly. GenCat currently includes Input, Edit, Browse, and Search functions, and offers access to a range of the repository contents through REST-style Web services, as well as consuming external SOAP-based Web services. Similarly, the LTER new *metadata editor project* (Aguilar et al., 2008) provides a full-fledged Web-based editor and entry form tool for metadata. The required future features of this editor will include: compliance with multiple metadata standards [such as EML and Federal Geographic Standard Committee (FGDC) and its Biological Data Profile or (BDP)], autosave feature, content validation, examples, metadata best-practices guidance, authentication system, integration of thesauri and taxonomic Web services, and Google maps mashups. Given the overlap and the use of the same technologies, the LTER and GSC have agreed on exploring the possibility of joining forces and develop either only one tool, or leverage common modules and expertise the projects. LTER also develops crosswalks (a mapping correspondence) between its and metadata standards used by other scientific communities. LTER is exploring a possible crosswalk between the GCS and LTER metadata standards, and some possible options were highlighted at the workshop. Because at this point both implementations are not tuned sufficiently, it seems wiser to wait until both standards evolve to favor a more seamless integration than in its current state.

Roadmap and Achievements

The future: actions and recommendations

The LTER Network endorses the activities of the GSC and will make the following specific contributions to this community. It will help

- support the outreach efforts of the GSC by advocating MIGS/MIMS adoption within the LTERs

- provide GSC guidance on the development and implementation of the GCDML
- recommend enhancements to EML to support genomic data,
- explore LTER collaboration with the GSC on training and;
- continue efforts to build on the similarities between the GSC's GenCat and the LTER's Metadata Editor projects.

The LTER will also increase efforts to reach out to researchers generating genomes and metagenomes. A fundamental ongoing activity initiated to support these goals is an in-depth evaluation of how many LTER research projects involve sequencing genomes. The LTER microbial ecology program that began in 1999 serves as an excellent case study (Microbial LTER Project, 1999). LTER principal investigators associated with the microbial ecology program are being asked to provide some guidance and their own research plans as examples of LTER interest in genomic research. Once we receive feedback from all the LTER microbial projects as well as from others, we will propose to the LTER Information Managers Executive Committee (IMExec) and LTER Network Information System Advisory Committee (NISAC) an extended work plan that goes beyond the collaboration items laid out in this report. NISAC guides the work performed by the Network Information System personnel, a substantial part of the LTER network wide informatics resources.

The LTER also propose draft designs for the extension of EML. The ecological community with the GSC would greatly benefit from using parts of the EML schema to document their reports, as there is substantial support and knowledge in the community. However, EML lacks adequate information placeholders to properly describe genomic and metagenomic data. There are several possibilities for enhancing the EML schema to accommodate this data and benefit from GSC work in this area.

Three key options are being considered:

1. An additional module (incorporation of GCDML) could be proposed for genomic data, parallel to EML's four current modules. This top-hierarchy branch addition is particularly interesting from the point of view of technical schema merging aspects, as well as related tools to edit, ingest or use metadata in an analytical framework.
2. A second possibility would be to expand the existing EML dataset module to accommodate a new entity type for genome sequences, which for portability, would follow the GCDML core schema.
3. A third option could be implemented without changing the current EML schema at all, but instead, would make use of its built-in extensibility and internal references. A GCDML data description can be included as "additional metadata" and its content referenced by an "otherEntity." A similar method is currently used in EML to refer to measurement units described using the Scientific, Technical, and Medical Markup Language (STMML) (Murray-Rust and Rzepa, 2002) schema.

The last option would result in fully compliant EML that contained descriptions of genomic data; however, any machine interpretation would require that standardized practices first be established. We argue that a full integration of the GCDML schema into EML (option 1) is the best long term solution.

The future: integration of ecological and genomic/metagenomic data with other data types

Data integration must extend beyond considerations involving EML and GCDML. There is a clear need in the LTER community for sharing microbial diversity data, whether sequence, metagenome or genome-based. There are many LTER participants doing broad-scale microbial studies, discussed more below, that include a wide range of research outputs in a wide variety of measurement and file formats. Imagine now comparing across these multiple studies; doing so without standardization would prove an exceedingly difficult task. As opposed to forcing researchers into making *a posteriori* comparisons, why not solve the problem from the ground up? The following case studies provide a view on the needs and complexity of LTER-associated microbial diversity projects.

The first case study is Linda Kinkel's (2008) broad-scale microbial studies, which include sequence data, details of organism isolation procedures, HPLC, mass spec profiles, nutrient utilization data, and other phenotypic data. Linda's group is considering a number of data integration tools that would enable them to integrate diverse types of data (such as images, matrices of quantitative and qualitative information, sequences, etc.). The North Temperate Lakes (NTL) LTER site provides another case study that identifies the needs to link environmental datasets and genomics databases (Jacob et al., 2007). The NTL-Microbial Organisms (MO) Environmental Sequence Database effectively link microbial community composition and environmental data from north temperate lake ecosystems. NTL-MO dataset includes community fingerprints, 16S rDNA sequences, phylogenetic assignments, and environmental data collected using high-throughput techniques. The data are stored in a relational database that is fully integrated with the North Temperate Lakes LTER datasets, linking molecular data with ecological data. The NTL microbial laboratory need to develop their own local system to manage the geospatial nature of their microbial genomics research places the lack of an adequate metadata standard in the spotlight. EML, as the LTER Network standard would have to be enhanced to address the work conducted at NTL-MO's microbial research. We feel confident that with an integration of a mature GCDML, the data and projects conducted at NTL-MO will be adequately documented. NTL-MO's McMahon, with the support of her coworkers, has pledged help in developing adequate metadata standards as outlined in this report.

Another case study example is the Alpine Microbial Observatory (AMO), which is associated with the Niwot Ridge LTER. Research at the alpine microbial observatory focuses on studying the diversity and function of soil micro-organisms across extreme environmental gradients in alpine ecosystems. In order to address these questions, AMO researchers take samples from different locations and collect soil biogeochemistry and sequence data, all of which is stored in a locally developed database. The AMO database carries the x, y, z, and t measurements that are core to MIGS/MIMS (Field et al., 2008), and therefore also GCDML

(Kottmann et al., 2008). The sequence and environmental data structures in the MIGS/MIMS specification, and therefore the GCDML schema, have their parallels in AMO database tables—the attributes of georeference, sampling event, sequence, and environmental data can be mapped between AMO and the metadata standards discussed in this paper. Among the elements that AMO's workflow has required are basic technical attributes of data such as descriptors, quantities, units of measure, and error estimates. In addition, methodology, administration, and attribution information is also present in the AMO workflow and research priorities. The AMO database will soon be able to import, export, and query MIGS/MIMS compliant data in GCDML.

Perhaps the best representative case study that reflects the value of the standards and tools sponsored by the GSC community is Linda Amaral's Microbial Inventory Research Across Diverse Aquatic (MIRADA) project (Amaral, 2008). This project proposes to establish a Microbial Biodiversity Survey and Inventory across all the major aquatic (marine and freshwater) LTER sites. MIRADA's proposed Biodiversity Survey and Inventory takes advantage of the aquatic sampling locations that are part of the established LTER network of sites and builds on existing infrastructure for coordination at the Marine Biological Laboratory (MBL) in Woods Hole, Massachussets, under the project International Census of Marine Microbes (ICoMM). MIRADA will adopt ICoMM's massively parallel, 454-based rDNA tag sequencing strategy that allows extensive sampling of both common and rare members of microbial populations, and provides a common metric for integrating studies of microbial diversity across aquatic LTER sites. This strategy, based on sequencing of hypervariable regions of the small subunit ribosomal RNA gene, has the ability to recover large sample sizes (100 to 1000 times the amount of information recovered from a typical clone library survey approach) of all components of the microbial community—*Bacteria*, *Archaea*, and *Eukarya*. This will not only enable cross-site comparisons, but also provide valuable baseline data for integrating population structures with ecosystem change, and understanding microbially mediated trophic dynamics and biogeochemical processes—areas of study already underway at many of the LTERs. Amaral's specific objectives are to:

1. Document and describe both microbial (*Bacteria*, *Archaea*, and *Eukarya*) baseline diversity and relative abundance data for microbial operational taxonomic units (OTUs) as defined by SSU rDNA hypervariable tags at aquatic LTER sites.
2. Determine which microbial OTUs are common to both freshwater and marine LTERs.
3. Determine whether diverse aquatic LTER sites possess "signature" assemblages characterized by space, time, and environmental parameters.
4. Discover novel tag sequences that likely represent novel micro-organisms that LTER researchers and students can further characterize and study.

Some of the projected MIRADA LTERs program measurable impact includes the publication of primary data by the participating LTER partners, and release of data in a variety of formats for the wider community.

The MIRADA project involves designing and developing a custom metadata and data repository for this cross-synthesis project. The MIRADA participants will investigate a way to bridge the content details for the MICROBIS database and the MIGS/MIMS, GCDML, and/or EML reports.

Wade Sheldon developed a comprehensive relational database for managing environmental sequence data and metadata at the Sapelo Island Microbial Observatory (SIMO) in 2001 (Sheldon and Moran, 2001; Sheldon et al., 2002). The database schema was designed to reflect the natural hierarchy that exists among samples and information and materials derived from them and to maintain the complete research context for data stored in the system, including environmental context of samples, field and laboratory methodology, and analytical post-processing. Interactive web applications support querying the database by both environmental and taxonomic characteristics, and retrieving results in standard bioinformatics file formats for analysis in other systems. The SIMO database is also coupled to automated bioinformatics pipelines for classification of 16S rRNA sequences and submission of annotated sequence data to the NCBI GenBank database, including all environmental and research context modifiers supported by NCBI. Sequence records and metadata can be retrieved using either the SIMO ID or GenBank Accession, and GenBank links are displayed on sequence detail pages to support bi-directional queries between both systems. A public version of the SIMO database has also been developed to provide access to the SIMO 16S rRNA classification pipeline and GenBank submission tool, including basic metadata forms for entering environmental, geographic and research context descriptors to accompany the sequence data set.

The success of the SIMO environmental sequence database has inspired similar efforts by other NSF-funded Microbial Observatories, including the North Temperate Lakes MO, Red Layer MO, and Alpine MO. The SIMO database design was also influential in the development of the International Census of Marine Microbes MICROBIS database (Neal et al., 2006). The SIMO database could potentially serve as a prototype for web-based systems that provide useful analytical services as part of the data submission process, and support dynamic generation of GCDML-compliant metadata as a value-added product.

Finally, note that there are many more case studies at LTER sites that we would have included in this short report, and such supplemental information can be found at the GSC WiKi and links therein.

Final message and specific actions

LTER is committed to a number of action items that will bridge the technical gaps it currently faces in capturing (meta) genomic data. By collaborating with GSC, LTER will be able to leverage the efforts in designing a comprehensive metadata standard for genomic and metagenomic data that is currently needed by the ecological genomics community. In addition, LTER brings experience to the table on successfully championing a community standard. We see LTER as a stakeholder in this wide community effort initiated by the GSC, and we feel the importance of being involved in a process that will directly affect how the LTER community plans and conducts site and network-oriented research.

Author Disclosure Statement

The authors declare that no competing financial interests exist.

References

Aguilar, R., Gries, C., and San Gil, I. (2008). LTER Metadata Editor Project. (http://intranet.lternet.edu/im/project/MetadataEditor).

Amaral, L. (2008). (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0717390).

Field, D., Morrison, N., et al. (2006). Meeting Report: eGenomics: Cataloguing our Complete Genome Collection II. OMICS **10,** 100–104.

Field, D., et al. (2008). Towards a richer description of our complete collection of genomes and metagenomes: the "Minimal Information about a Genome Sequence." Nat Biotechnol **26,** 541–547.

Jacob, C., Brent, A.D., Benson, B.J., Newton, R.J., and McMahon, K.D. (2007). Proc. 9th World Multiconference on Systematics, Cybernetics and Informatics.

Jones, M., et al. (2001). EML—http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html

Kinkel, L. (2008). Spatial Variation, Diversity and Genetic Composition of Microbes in Prairie Soils (http://www.cedarcreek.umn.edu/microbo/index.htm).

Kottmann, R., Gray, T., et al. (2008). A standard MIGS/MIMS compliant XML Schema: Towards the development of the Genomic Contextual Data Markup Language (GCDML). OMICS (this issue).

LTER Information Managers Executive Committee (2005). http://intranet.lternet.edu/im/im_requirements/im_review_criteria

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G. (1997). Nongeospatial metadata for the ecological sciences. Ecol Appl **7,** 330–342.

Microbial LTER Project. (1999). (http://www.lternet.edu/microbial_ecology/).

Murray-Rust, P., and Rzepa, H.S. (2002). STMML. A markup language for scientific, technical, and medical publishing. Data Sci J **1,** 128–192.

San Gil, I., and Baker, K. (2007). Databits. http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/07fall/#fa2

Sheldon, W.M., Moran, M.A., and Hollibaugh, J.T. (2002). Efforts to link ecological metadata with bacterial gene sequences at the Sapelo Island Microbial Observatory. Pages 402–407 in: Proceedings of the 6th World Multiconference on Systemics, Cybernetics, and Informatics. Information Systems Development II. International Institute of Informatics and Systemics, Orlando, Florida.

Sheldon, W.M., and Moran, M.A. (2001). Sapelo Island Microbial Observatory Environmental Sequence Database (http://simo.marsci.uga.edu/public_db/).

Neal, P., Patterson, D., and Bordenstein, S. (2006). MICROBIS: The ICoMM Marine Microbes Database (http://icomm.mbl.edu/microbis/).

Address reprint requests to:
Inigo San Gil
LTER Network Office
University of New Mexico
MSC03 2020
Albuquerque, NM 87131

E-mail: isangil@lternet.edu

VIII. Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata

Authors: Lynette Hirschman, Cheryl Clark, K. Bretonnel Cohen, Scott Mardis, Joanne Luciano, Renzo Kottmann, James Cole, et al.

Published in *Omics*. 2008; 12(2): 129-36.

Contribution: Term definition, result check, use case definition.

OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008

DOI: 10.1089/omi.2008.0016

Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata

Lynette Hirschman,[1] Cheryl Clark,[1] K. Bretonnel Cohen,[1] Scott Mardis,[1] Joanne Luciano,[1] Renzo Kottmann,[2] James Cole,[3] Victor Markowitz,[4] Nikos Kyrpides,[5] Norman Morrison,[6] Lynn M. Schriml,[7] Dawn Field,[8] and the Novo Project[9]

Abstract

There is an urgent need to capture metadata on the rapidly growing number of genomic, metagenomic and related sequences, such as 16S ribosomal genes. This need is a major focus within the Genomic Standards Consortium (GSC), and *Habitat* is a key metadata descriptor in the proposed "Minimum Information about a Genome Sequence" (MIGS) specification. The goal of the work described here is to provide a light-weight, easy-to-use (small) set of terms ("Habitat-Lite") that captures high-level information about habitat while preserving a mapping to the recently launched Environment Ontology (EnvO). Our motivation for building Habitat-Lite is to meet the needs of multiple users, such as annotators curating these data, database providers hosting the data, and biologists and bioinformaticians alike who need to search and employ such data in comparative analyses. Here, we report a case study based on semiautomated identification of terms from GenBank and GOLD. We estimate that the terms in the initial version of Habitat-Lite would provide useful labels for over 60% of the kinds of information found in the GenBank isolation_source field, and around 85% of the terms in the GOLD habitat field. We present a revised version of Habitat-Lite defined within the EnvO Environmental Ontology through a new category, EnvO-Lite-GSC. We invite the community's feedback on its further development to provide a minimum list of terms to capture high-level habitat information and to provide classification bins needed for future studies.

Introduction

THIS PAPER DISCUSSES the current status of an ongoing effort to create a minimum hierarchical controlled vocabulary for the capture of habitat and environmental metadata on genomics, metagenomics, and 16S ribosomal sequences. This work has two goals. The short-term goal is to develop a light-weight controlled vocabulary (Habitat-Lite) within the EvnO framework to capture high-level habitat and environmental metadata in support of the Genomic Standards Consortium (GSC) Minimal Information about Genome/Metagenome Sequence (MIGS/MIMS) specification (Field et al., 2008a, 2008b). The longer-term goal is to develop a repeatable process for other types of metadata by identifying key terms based on usage in databases and the open literature. We will evaluate the coverage, utility, and usability of the key terms and refine the set of terms based on these measures. Additionally, we will develop tools to facilitate the capture of the metadata from free text fields.

This effort originated in the context of the development of the MIGS/MIMS checklist (http://gensc.org/gc_wiki/index.php/MIGS/MIMS), and has also been discussed in the context of the newly established Environment Ontology (EnvO) project (http://environmentontology.org—see the GSC EnvO wiki page for ongoing discussion: http://gensc.org/gc_wiki/index.php/EnvO_Project; also see the EnvO SourceForge site: http://obo.cvs.sourceforge.net/obo/obo/ontology/environmental/), as part of advocating the use of ontologies

[1]Information Technology Center, The MITRE Corporation, Bedford, Massachusetts.
[2]Microbial Genomics Group, Max Planck Institute for Marine Microbiology and Jacobs University Bremen, 28359 Bremen, Germany.
[3]Center For Microbial Ecology, Michigan State University, East Lansing, Michigan.
[4]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California.
[5]Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California.
[6]School of Computer Science, University of Manchester, Oxford Road, Manchester, United Kingdom.
[7]Institute for Genome Sciences and Department of Epidemiology and Preventive Medicine, University of Maryland School of Medicine, HSFI, 685 West Baltimore Street, Baltimore, Maryland.
[8]NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford, Oxfordshire, United Kingdom.
[9]http://environmentontology.org

in capturing MIGS/MIMS reports. See also GazO, http://gensc.org/gc_wiki/index.php/GAZ_Project, a first step towards an open source gazetteer, constructed on ontological principles, that describes places and place names and the relations between them. This work is informing GSC consensus-building activities, and has led to agreement to adopt the Habitat-Lite terminology for use in the Genomic Contextual Data Markup Language (GCDML) (Kottmann et al., 2008).

There is a strong need for developing methods to facilitate the capture of metadata describing the growing number of genomic and metagenomic projects, including information about isolation source and habitat (Field et al., 2008a; Morrison et al., 2006). The increase in the associated literature is also accelerating, particularly in light of projects such as the Global Ocean Survey (Venter et al., 2004) and the Human MicroBiome (http://nihroadmap.nih.gov/hmp/) (Gill et al., 2006), with parallel growth in the relevant databases [see, e.g., Fig. 1 of Morrison et al., 2006, for an illustration of exponential growth in the number of sequences in the International Nucleotide Sequences Database Collaboration (INSDC)]. However, the capture of the metadata associated with these projects remains a major challenge, largely due to the fact that the literature is scattered and the metadata is difficult to find, even by expert manual extraction. Many databases have fields to support the capture of metadata, but such entries are often sparse and are entered as free text, thus lacking standardization in vocabulary and definitions, impeding our ability to perform meaningful comparisons or utilize information from multiple resources. The case studies discussed below illustrate the resulting difficulty in using computational techniques to study the relation between habitat and genotypic or phenotypic properties of organisms (Hunter, 2002, von Mering, 2007)—a key goal of genomic and metagenomic studies.

Our initial work has focused on a specific metadata type, namely habitat. For our purposes here, we define habitat as "the place or environment where an organism naturally or normally lives and grows." It is distinguished from "sample source," which is the environmental context in which a sample is collected, as defined in Morrison et al. (2006). Multiple habitat terms can be associated with a species; by contrast, a sample is associated with a description of its (unique) source. Table 1 shows excerpts from the GOLD database (Liolios et al., 2008); we can see that the "Habitat" field often has multiple entries, in contrast to the "Isolation" field, which describes the specific sample source and is much more detailed. The initial version of Habitat-Lite is aimed at capturing high-level habitat descriptions; ongoing work on the environmental ontology EnvO will provide a much finer-grained set of terms to describe specific environments and sample source information.

The development of Habitat-Lite began with the selection of a small list of widely used high-level terms for describing habitat. We used these terms to "bin" information contained in free text fields for habitat or source information in several key databases. This process enables us to develop measures of coverage, utility, and usability for the term set—for example, how well the controlled vocabulary covers the free text entries, how evenly the entries are distributed across the bins defined by the controlled vocabulary, how well the bins capture useful categories for search, how cost-effectively the controlled vocabulary terms can be used to annotate new data, and how consistent the mappings are across multiple annotators (human or automated). There are trade-offs in this complex space between the detailed information that can be captured with a large well-structured set of terms (e.g., an ontology), versus the time it takes to create a stable set of structures and the cost of acquiring consistent annotation using this much richer terminology, including supporting tools.

The two major data sources chosen for this study contain large numbers of records and descriptors of habitat in free text form. Ideally, we would have looked in the literature to

Table 1. Habitat and Isolation Fields from the GOLD Database

Organism	*Strain*	*Phenotype*	*Habitat*	*Isolation*
Hemophilus influenzae NTHi	PittEE	Pathogen, facultative, nonmotile, rod-shaped	Host	Middle-ear effusion of a child in Pittsburgh
Mycobacterium tuberculosis	H37Ra	Pathogen, aerobe, chemoorganotroph, rod-shaped, nonmotile	Host, TB epidemic	Original human-lung H37 isolate in 1934
Psychrobacter sp.	PRwf-1	Psychrophile, radiation resistant, rod-shaped, nonmotile	Aquatic, soil, Permafrost	
Roseiflexus sp.	RS-1	Filament-shaped, photosynthetic, thermophile, facultative, nonsporulating, motile, rod-shaped	Aquatic, hot spring	Hot spring microbial mat
Lactobacillus reuteri	F275 (JCM 1112)	Probiotic, non-pathogen, rod-shaped, facultative, nonmotile	Intestinal flora	Human isolate that is unable to colonize the intestinal tract of mice
Pseudomonas putida	F1	Aerobe, motile, rod-shaped, non-pathogen	Soil	Polluted creek in Urbana, Illinois, by enrichment culture with ethylbenzyne as a sole source of carbon and energy

determine how habitat and isolation source were described. However, for the initial experiments, it was much more efficient to look at fields in existing databases. The two sources were:

1. GenBank (http://www.ncbi.nlm.nih.gov/Genbank/): the *isolation_source* field, which captures free text descriptions, in the form entered by submitters to GenBank, related to sample source;
2. Genomes On-Line Database (GOLD) (http://www.genomesonline.org/) (Liolios et al., 2008): the *Habitat* field, which captures terms collected from the literature.

Development of Habitat-Lite

As a starting point, one author (D.F.) did a survey for terms used in a number of relevant sources. From this list, she selected a set of high-level terms as a strawman for the first iteration of the Habitat-Lite term list (shown in Table 2). The number of terms was kept small (less than 20), based on discussions with annotators at NCBI (we met with Tatiana Tatusova, Scott Federhen, Karen Clark, and Anji Johnston at NCBI Entrez Genomes, to explore ways to improve the capture of environmental/habitat metadata in GenBank), but could grow in future iterations. Our approach was to identify a set of seed terms, run experiments to determine how well these could "bin" existing entries, determine how useable such a set of terms would be for human and semiautomated annotation, and then iterate, with the goal of producing a consensus-driven "minimal set" of habitat terms that provided good coverage of entries in key resources. Both the original version and the new version of Habitat-Lite using the EnvO-Lite-GSC category are available in .obo format (http://gensc.org/gc_wiki/index.php/Habitat-Lite) and, for example, could be used with OBO-Edit (http://oboedit.org/), CoBrA (http://cobra.umbc.edu/eclipse/), or the Phenote a notation tool (available at http://www.phenote.org/).

The initial list of terms drew on previously published lists of habitat terms used to annotate databases (NCBI Microbial genomes; http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi), on proposed new community standards for the annotation of 16S sequences (http://www.jgi.doe.gov/16s/saiform.php), on the habitat terms published in the Global Ocean Survey (Nealson and Venter, 2007), on habitat terms used to describe the biases in culture collection strains (Floyd et al., 2005), and on patterns and biases in the complete genome collection (Martiny and Field, 2005); see Supplementary Table 1 for a full listing. These terms were mapped to an early version of the Environment Ontology (http://environmentontology.org), as shown in column 3 of Table 2.

Table 2. Initial Habitat-Lite Terms and Mappings to EnvO (Oct. 2007)

Habitat-Lite Terms for genomes and metagenomes		
1	freshwater habitat	ENVO:00000873
2	marine habitat	ENVO:00000447
3	terrestrial habitat	ENVO:00000446
4	soil	ENVO:00001998
5	water	ENVO:00002006
6	air	ENVO:00002005
7	sediment	ENVO:00002001
8	sludge	ENVO:00002044
9	waste water	ENVO:00002006
10	hot spring	ENVO:00000051
11	hydrothermal vent organism-associated	ENVO:00000215
12	habitat	ENVO:00002032
13	extreme habitat	ENVO:00002020
14	food	ENVO:00002002
15	biofilm	ENVO:00002034
16	microbial mat	ENVO:01000008
17	fossil	ENVO:00002164

Use Cases: Analyses Based on Habitat Data

Habitat-Lite terms were assembled from existing terminologies with the explicit goal of supporting as many use cases as possible—in particular, the ability to "bin" data into interesting categories for purposes of comparison. The use of bins is particularly attractive to biologists, who, for example, wish to extract sequences only associated with "soil bacteria" or "freshwater metagenomes." In this respect, biologists' descriptions of "habitat" contrast strongly with those of environmental scientists, who tend to describe habitat in terms of continuous variables.

We are now in the process of assembling use cases to test the coverage of Habitat-Lite. At the fifth GSC meeting, one author (J.C.) presented a small study done on the Ribosomal Database Project (RDP; http://rdp.cme.msu.edu/; Cole et al., 2007). The RDP consumes GenBank documents for 16S sequences and maintains them in a highly value-added format. These data are used extensively for contrastive analysis based on environmental factors. To determine both the coverage of environments and the utility of the habitat or environmental information in RDP, a small experiment was carried out in late 2006.

Using information from the INSDC records, one author (J.C.) attempted to manually classify into habitats the 168,911 rRNA sequences marked as *environmental* in RDP release 9.44 (November 2006). The habitat categories that were suggested by Phil Hugenholtz (DOE Joint Genome Institute) were modified by splitting *host-associated* into separate categories for *plant* and *animal* (including *human*) *associated*. We first assigned 24.5% of the sequences using their *isolation_source* qualifier. For those sequences without an isolation source tag, or where we were unable to classify based on that tag, we examined the reference titles from the INSDC records and were able to classify another 37.5% of the records. References used by fewer than 150 sequences were not examined because of the effort involved. The remaining 38% of sequences could not be classified because, for the most part, they did not have any habitat information in the INSDC record. Most assignments were made after examination by a single researcher, but spot-checking by a second researcher gave disagreement in assignment for only a small percentage of sequences. By far the biggest category was *animal associated*, and a large fraction of these were *human associated*. The *soil*, *sediment*, and *water* categories also represented large numbers of sequences (see Fig. 1).

A second interesting use case was reported in (von Mering et al., 2007). In this paper, the authors studied the association of preferred habitats for microbial clades and looked for correlations between evolutionary distance and similarity of habitat. The habitat information was taken from free

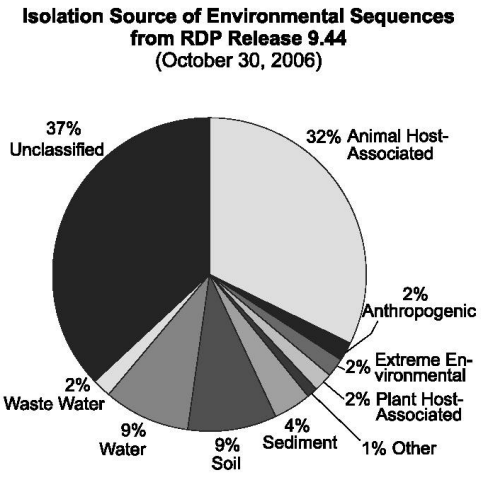

FIG. 1. Categorization of isolation source for environmental sequences from RDP.

text fields of the Greengenes database (Desantis et al., 2006) and the microbial culture collections (Dawyndt et al., 2005). To assess similarity of habitat, the authors manually selected "informative" words found in the annotation of five or more experiments (von Mering et al., 2007, Tables S2 and S3, supplemental materials) and computed a pairwise similarity score between habitats, based on number of shared keywords. Graphs in Table S2 Figure 2B and 2C show that more habitat "features" are shared among the more closely related organisms, both in terms of taxonomy and molecular similarity.

These use cases illustrate the kinds of information that would be useful to researchers, and also the difficulties of obtaining the information in the absence of a common underlying controlled vocabulary.

GenBank "isolation_source" Entries

To validate and refine the selection of Habitat-Lite terms, examples of habitat or isolation_source information were needed to determine what information was present, and how this information was expressed. An ideal approach would be to extract metadata from the published literature; however, this is quite difficult, because the metadata occurs in many diverse forms, including PDF tables, densely written materials and methods sections, supplementary material, and even in referenced work. Therefore, we took advantage of the large quantity of free text metadata already available—as fields in database records. As a first step, we analyzed the "isolation_source" field from GenBank gene records, which captures, as short free text entries, information about the isolation source of the specific sequence being deposited. John Wilbur (NCBI) provided us with a list of 35,000 distinct isolation_source entries from GenBank gene records as of September 2007—see Table 3 for examples of some entries from

TABLE 3. DISTRIBUTION OF UNIQUE HABITAT-LITE TERMS IN GENBANK *ISOLATION SOURCE* FIELDS

Class	*Table frequency*	*Data set frequency*	*Percent total*	*Example*
Organism-associated	14,781	341,003	42.4%	1-year-old male spleen
Water/aquatic	2,008	40,794	5.8%	0 m water at a station in the North Atlantic
Soil	1,115	229,032	3.2%	0–20-cm bulk soil from a mixed forest
Marine	944	3,115,879	2.7%	0.2–0.8 μm fraction from surface sea water
Sediment	723	34,435	2.1%	aquaculture coastal sediments
Terrestrial	594	3,100,550	1.7%	a declining forest
Food	398	4,003	1.1%	(onion)
Sludge	294	9,868	0.8%	1st maturation stage of sludge
Microbial mat	195	9,164	0.6%	a deep sea microbial mat
Waste water	195	5,969	0.6%	activated tannery effluent from treatment plant
Hydrothermal vent	133	3,036	0.4%	14 N Mid Atlantic Ridge Logatchev vent field
Hot spring	121	3,249	0.3%	6–48 celsius region of a hot spring
Extreme	117	4,967	0.3%	a solar saltern
Biofilm	114	3,499	0.3%	aquatic phototrophic biofilm
Freshwater	75	2,609	0.2%	Arctic freshwater lake
Fossil	67	507	0.2%	100,000-year-old fossil
Air	21	768	0.1%	African air sample
Total Habitat-Lite terms	21,896		62.9%	
Total unique terms	34,836			

this field. We were primarily interested in whole genome or metagenomics sequences, but the initial data set consisted of entries for all genes. As a result, frequency counts were heavily skewed towards large metagenomics projects, so we did not use the frequency counts in our analysis. For example, the phrase "locations in the Sargasso Sea, Panama Canal, and the Galapagos Islands" occurred over 3 million times in this data set.

Because of the size of the data set, it was not possible to explore it manually. One of the authors (C.C.) developed a small set of scripts to identify probable classes based on the presence of specific key words in each entry. The key words used for this analysis were based on the original Habitat-Lite terms plus synonyms and, in some cases, specializations. For example, for "waste water" the terms used for matching were "waste water," "waste-water," "wastewater," "sewage," "sewerage," etc. For "food," the terms used for matching included specific kinds of foods, for example, "milk," "cheese," "beer," etc. Similarly, for "organism-associated," the terms used for matching had to capture the many ways of expressing specific organisms, particularly humans, for example, "patient," "female," "subject," "child," etc.

Of the almost 35,000 distinct entries in the isolation_source field, some 22,000 (63%) contained specific words or phrases that could be mapped to the 17 Habitat-Lite categories. The bulk of these fell into the Organism-associated category (42%). In addition, we were able to identify over 20% of the entries that were geographic names or temporal expressions or other numerical quantities or identifiers. This enabled us to account for approximately 85% of the entries from isolation_source. The remaining 15% contain low frequency terms—many of them with species information ("wild mulberry"), location information ("Wilson and the Australian Museum)," or information about culture techniques ("top band of HTA gel").

This pattern-matching approach allowed us to obtain a quick overview of the types of information found in the GenBank isolation_source field. This approach would require significant refinement and/or human intervention if we wished to use it for semiautomated assignment of Habitat-Lite terms to isolation_source entries, for improved search and indexing. In particular, this strategy mapped each entry to a single field, so that, for example, *130 m below sea surface* was mapped only to "marine," losing the depth information. Similarly, the entry "Marine Biology Laboratory" caused the entry to be associated with the category "Marine"—a plausible inference but certainly not explicit information about habitat.

Habitat Field Entries from the GOLD Database

We next investigated a second data set, which consisted of the Habitat entries from the GOLD database on October 2007. The initial data set consisted of 1455 entries with 2210 terms. Table 1 shows some example GOLD entries, including not only the Habitat field, but also the much more detailed Isolation field. The entries in the Habitat field frequently contained multiple entries that specified the range of known habitats for a specific organism, for example, "Host, TB epidemic" or "Aquatic, Soil, Permafrost."

Coverage of GOLD terms using Habitat-Lite

First, we looked for exact matches between GOLD Habitat terms and Habitat-Lite terms plus the additional term "aquatic." This resulted in exact matches for 84% of GOLD Habitat terms. The three most frequent terms ("host," "aquatic," and "soil") covered 75% of GOLD habitat data, while six Habitat-Lite terms were not seen at all in this smaller data set ("air," "freshwater," "extreme," "microbial mat," "fossil," "terrestrial").

Comparison of automated mapping and expert mapping

In the next experiment, we applied the automated mapping used in the GenBank experiment to the unique entries in the GOLD Habitat data, and compared these results to an expert mapping done by one of the authors (D.F.). There were a total of 132 unique entries in the GOLD Habitat field for metagenomes. There was 64% agreement (84 of 132) and 48 cases of differences in the automated mapping versus expert mapping. Most differences were due to a failure in the automated mapping procedure (30 cases, which were not classified or not mapped to the limited controlled vocabulary). Another nine were due to mismatches related to the new category "aquatic" introduced by the expert (five) and four were due to difficulty in classifying between freshwater and water.

The remaining nine discrepancies (shown in Table 4) brought to light interesting problems. Several of the discrepancies pointed out an ambiguity in the classification scheme with respect to "extreme habitat:" terms such as "hot springs," "permafrost," and "hypersaline mats" could be classified as "extreme habitat" or into a geographic or environmental feature ("hot springs," "soil," "microbial mat"). In another case ("rice paddies"), it is unclear without further context whether the focus was on the *rice* in rice paddies ("organism-associated") or on the *paddies* ("terrestrial").

TABLE 4. EXAMPLES OF DISAGREEMENT BETWEEN EXPERT MAPPING AND AUTOMATED MAPPING FOR THE GOLD HABITAT DATA

GOLD Habitat term	*Expert mapping*	*Automated mapping*
Mud	terrestrial	Soil
Rice paddies	terrestrial	Organism-associated
Soda lakes	organism-associated	Water
Hot spring	extreme habitat	Hot spring
Permafrost	extreme habitat	Soil
Snow	extreme habitat	Freshwater
Sulfur spring	extreme habitat	Water
Hypersaline mats	microbial mat	Extreme

These examples illustrate well the need for annotation guidelines, to handle situations where a term might be placed in several categories. There are several possible solutions: either there need to be "orthogonal dimensions" that would allow a category like "extreme habitat" to be "checked off" separately from some more specific information about geographic or environmental features or alternatively, there could be a facility to allow a given term to belong to multiple "bins."

Manual annotation of the GOLD data to two orthogonal bins

The final set of experiments was designed to test the difficulty of the annotation task and to determine whether better annotation could be done by assigning multiple orthogonal terms. As noted above, there are advantages to capturing orthogonal annotations: to preserve richer information for searching, and also to reduce interannotator disagreement. To experiment with this approach, a single author (K.B.C.) annotated the 132 GOLD unique terms using Habitat-Lite in conjunction with an explicit set of guidelines that were meant to ensure that every *Habitat* entry was assigned both a general (*biome*) term and an *environment* term. The guidelines made use of the mappings of the Habitat-Lite terms to the EnvO taxonomy as follows:

1. Assign a child term of *biome* (*freshwater*, *marine*, or *terrestrial*).
2. Can the input be assigned a child class of *habitat* (*organism-associated* or *extreme*)? If so, assign it, and then stop. (This had an undesired effect, which we describe below.)
3. Is the input a food? If so, assign *food*. If not, go to (4).
4. Can the input be assigned a child of *biotic/abiotic* (*biofilms*, *microbial mat*, or *fossil*)? If so, assign it, and then stop. If not, go to (5).
5. Can the input be assigned a child class of *hydrographical/physiographic/anthropogenic* (*hot spring*, *hydrothermal vent*, or *wastewater*)? If so, assign it, and then stop. If not, go to (6).
6. Can the input be assigned a child of *environmental substance* (*soil*, *water*, *sediment*, *sludge*, or *air*)? If so, then assign it.
7. Stop.

The undesired effect of Step (2) was that some inputs that could have been assigned specific terms related to extreme habitats were instead only assigned the more general *extreme (habitat)*. A simple reordering of the rule might fix this.

The results demonstrate that the annotation task is well within the range of someone with reasonable background in biology. Only 2 out of 132 entries were left unannotated due to lack of domain knowledge: *solfataric fields*, and *self-heated organic materials*. It took approximately 1.5 h to do about 2 * 132 annotations, or around 1.5 terms annotated/minute. Based on this estimate, it would take less than a day's work to map all of the GOLD Habitat entries to Habitat-Lite.

Discussion

The goals of this work were to create a useful set of high-level terms to capture habitat data, and to develop a methodology that can be applied to similar problems—specifically, to:

1. Determine what descriptors of habitat are recorded and how they are expressed in free text;
2. Determine how well a small set of terms, such as Habitat-Lite, could cover terms found in key resources;
3. Examine the feasibility of (semi-)automated capture of the these fields of information for future projects.

Our initial experiments have resulted in a new version of Habitat-Lite (shown in Table 5), based on analysis of the GenBank *isolation_source* field and the *habitat* field in the GOLD

Table 5. Proposed Habitat-Lite Version 0.2

Top level	*Second level*	*Third level*	*Example also could be coded for*
Choose one or more:	Choose one or more:	Free text description, e.g.	
Aquatic: freshwater	soil	pinyon–juniper forest soil	Terrestrial
Aquatic: marine	sediment	oxygen-depleted intertidal marine sediment	Aquatic: marine
Aquatic	sludge	Thermophilic methanogenic sludge	Terrestrial?
Terrestrial	waste water	waste water of paper machine	Aquatic
Air	hot spring	hot spring at 70°C	Aquatic, Extreme
Fossil	hydrothermal vent	the shallow hot vent in Iwo Jima	Aquatic: marine, Extreme
	biofilm	biofilm of drinking water distribution system	Aquatic
	microbial mat	hot spring microbial mat	Aquatic, hot spring
Food	[Food Ontology or CV]	surface of smear ripened cheese	Food
Organism Associated	[Species CV] [Anatomy CV, e.g., MIAA]	gut of nitidulid beetle	Organism-Associated
Extreme habitat	Select if appropriate	extremely alkaline (pH 12 to 13) groundwater	Aquatic; Extreme
Other		(45.32739 N, 80.40874 W)	

database. Based on this analysis, we put forward the following recommendations for Habitat-Lite:

- A shift from a "flat" list to one with some structure is necessary.
- The set of terms should support certain inferences useful for search; for example, that a sample labeled *soil* is also *terrestrial*, or that a sample from a *hydrothermal vent* is also *extreme*.
- Consistent annotation requires guidelines for general terms such as *terrestrial* and *aquatic*, to instruct annotators to annotate to the most specific term possible.
- The notion of *extreme environment* is problematic in that it should be annotated **in addition** to a more specific term, such as *hot spring*—thus requiring that certain entries be associated with two Habitat-Lite terms.
- The category *Organism-associated* needs to be subdivided by linking out to other ontologies or controlled vocabularies (specifically, a taxon hierarchy and perhaps a high-level anatomy ontology).
- *Fossil* is an example of a currently infrequently used term, but a candidate for inclusion as a term of "exceptional importance" that could be useful in the future for searching.

The new version of Habitat-Lite has been implemented through the use of the Category mechanism in OBO-Edit with the category "EnvO-Lite-GSC." We use the name EnvO-Lite-GSC to indicate that the ontology is a "light weight" version based on the full EnvO but tailored to the needs of the Genome Standards Consortium. Other groups that need a specific view of EnvO can use this mechanism, which makes it possible to save a specific view of the full EnvO ontology (available from http://obo.cvs.sourceforge.net/obo/obo/ontology/environmental/), drawing on the EnvO identifier space and using same structure of the ontology and the same terms but including only the terms specified by the selected category. Such views can be requested from the EnvO curatorial staff via the EnvO-tracker or the desired view of EnvO can be created locally using OBO-Edit.

The new set of Habitat-Lite terms is structured into two levels: a set of high level terms (first column in the table: *aquatic*, *terrestrial*, *air*, plus *organism-associated*, *food*, *extreme environment*, *fossil*), and a second level of more specific terms (column 2 in Table 5). To maximize capture of information, this version encourages selection of one or more of the high-level terms, one or more of the second-level terms, and recording of the specific information in free text (column 3, Table 5). The free text is shown associated with its level 2 term and in column 4, one or more appropriate top level terms.

To maintain simplicity, there is no obligatory connection/restriction between choice of top-level terms and second-level terms, except for the "food" and "organism-associated" classes. This allows flexibility (e.g., there are both freshwater and salt marshes) with the downside of increased possibility for error or for incomplete annotation. It should be possible to do automated association of high level terms, based on the second level terms, for example, associating "terrestrial" automatically with any annotation of "soil," "sediment," or possible new terms such as "sand," "wood," "rock," or "mud."

The "organism-associated" class should be elaborated by a term describing the organism and an anatomy term for the part of the organism; we will investigate use of a minimal anatomy ontology, such as Jonathan Bard's MIAA (Minimal Information about Anatomy; personal communication). The food class for now is just left as free text; it may be possible to use a small specialized food controlled vocabulary or ontology in the future (see http://gensc.org/gc_wiki/index.php/Food_Ontology_Project for discussions about the creation of a food ontology or controlled vocabulary).

Next Steps for Habitat-Lite: Adoption by GOLD, RDP, GCDML

The new version of Habitat-Lite will be tested against the GOLD data and revised to support GOLD (Liolios et al., 2008), IMG (Markowitz et al., 2008a), and IMG/M (Markowitz et al., 2008b). (For IMG, see http://img.jgi.doe.gov; for IMG/M, see http://imgweb.jgi.psf.org/cgi-bin/m/main. cgi.) GOLD has embraced the adoption of this controlled vocabulary/ontology for its habitat data. Capture of GOLD and IMG habitat data is currently implemented via the Expert Review Web submission form on the Integrated Microbial Genomes (IMG) Web site. All genomes submitted directly into IMG and IMG/M are now required to provide metadata that conforms to the GOLD vocabulary. The RDP (Cole et al., 2007) has also agreed to adopt the revised version of Habitat-Lite. The new version of Habitat-Lite will be supported in GCDML (Kottmann et al., 2008).

Conclusions

These results indicate that it should be possible to produce a list of terms with good high-level coverage for Habitat-Lite. We accept that candidate Habitat-Lite terms provide only very high-level information, and that these terms may be an amalgamation of terms found in different branches of future ontologies, or even among different orthogonal ontologies (e.g., for "organism-associated"). We also recognize that while these terms may provide a useful tool for biologists and databases, they have severe limitations. We emphasize the importance of maximum reporting of information about habitat—in particular, the necessity of preserving free text fields associated with legacy data so that more fine-grained information is never lost, and reanalysis is always possible.

In the long term, our goal is the creation of an interactive metadata checking system (a kind of metadata "spell checker") that could "read" free text and suggest the correct mapping into a controlled vocabulary/ontology, for user validation or correction, thus ensuring that metadata is comprehensively captured and "binned" at the point of entry.

The use of a combination of Habitat-Lite terms in the short-term, cultural shifts in the way this community annotates to capture more complete descriptions of habitat and isolation source, and future use of ontologies and ontology-aware software will have a measurable benefit on the ability of researchers to effectively reuse ever-growing sources of data for large-scale, downstream analyses.

Toward a minimum information list of habitat terms for use in the GSC

We have posted the initial and revised versions of Habitat-Lite (Table 5) to the GSC Wiki. This list is annotated with recommendations and issues that will be addressed in revising it. We are making an open call for evaluation of this

list of habitat terms in order to develop a consensus-driven version of it that best suits community needs. This terms list will then be implemented in GCDML (Kottmann et al., 2008) and used in the first instance to fill the "Habitat" field of the MIGS-compliant Genome Catalogue database (http://gensc.org).

Acknowledgments

The work at MITRE (L.H., C.C., K.B.C., S.M., J.L.) has been supported in part by National Science Foundation Grant 0746650, Small Grant for Exploratory Research: Mining Metadata for Metagenomics. We thank Tatiana Tatusova for a number of discussions on an approach to developing a light-weight set of classes for annotation, as well as Scott Federhen, Karen Clark, and Anji Johnston of NCBI. We thank John Wilbur, NCBI, for providing us with the data from the GenBank isolation_source fields. We thank Norman Morrison and Lynn Schriml of the EnvO/Gaz project for critical reads of the manuscript.

Author Disclosure Statement

The authors declare that no competing financial interests exist.

References

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, S., McGarrell, A.M., et al. (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res **35,** D169–D172.

Dawyndt, P., Vancanneyt, M., Demeyer, H., and Swings, J. (2005). Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. IEEE Trans Knowledge Data Eng **17,** 1111–1126.

Desantis, T.Z., Hugenholtz, P., and Larsen, N. (2006). Greengenes, a Chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72,** 5069–5072.

Field, D., Garrity, G.M., Gray, T., Morrison, N., Selengut, J.D., Sterk, P., et al. (2008a). The Minimum Information about a Genome Sequence (MIGS) specification. Nat Biotechnol **26,** 541–547.

Field, D., Garrity, G.M., Sansone, S.-A., Sterk, P., Gray, T., Glockner, F.O. (2008b). Meeting Report: The 5th Genomic Standards Consortium Workshop. OMICS (this issue).

Floyd, M.M., Tang, J., Kane, M., and Emerson, D. (2005). Captured diversity in a culture collection: case study of the geographic and habitat distributions of environmental isolates held at the American Type Culture Collection. Appl Environ Microbiol **71,** 2813–2823.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Tumbaugh, P.J., Samuel, B.S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. Science **312,** 1355–1359.

Hunter, L. (2002). Ontologies for programs, not people. Genome Biol **3.**

Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., et al. (2008). A standard MIGS/MIMS compliant XML Schema: towards the development of the Genomic Contextual Data Markup Language (GCDML). OMICS (this issue).

Liolios, K., Mavrommatis, K., Tavernarakis, N., and Kyrpides, N.C. (2008). The Genomes OnLine Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res **36,** D475–D479.

Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I-M.A., et al. (2008a). The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res **36,** D528–D533.

Markowitz, V.M., Ivanova, N.N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., et al. (2008b). IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res **36,** D534–D538.

Martiny, J.B., and Field, D. (2005). Ecological perspectives on the sequenced genome collection. Ecol Lett **8,** 1334–1345.

Morrison, N., Wood, J.A., Hancock, D., Shah, S., Hakes, L., Gray, T., et al. (2006). Standard annotation of environmental OMICS data: application to the transcriptomics domain. OMICS **10,** 172–178.

Nealson, K.H., and Venter, J.C. (2007). Metagenomics and the Global Ocean Survey: what's in it for us, and why should we care? ISME **1,** 185–190.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science **304,** 66–74.

Von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., Jensen, L.J., et al. (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. Science **315,** 1126.

Address reprint requests to:
Lynette Hirschman
Information Technology Center
The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730

E-mail: lynette@mitre.org

IX. A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes

Authors: Peplies, Jörg, Renzo Kottmann, Wolfgang Ludwig, and Frank Oliver Glöckner

Published in *Systematic and Applied Microbiology*. 2008l 31, no. 4 (September): 251-257.

Contribution: Metadata definition.

Available online at www.sciencedirect.com
ScienceDirect
Systematic and Applied Microbiology 31 (2008) 251–257
SYSTEMATIC AND APPLIED MICROBIOLOGY
www.elsevier.de/syapm

A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes

Jörg Peplies[a], Renzo Kottmann[b,d], Wolfgang Ludwig[c], Frank Oliver Glöckner[b,d,*]

[a]*Ribocon GmbH, D-28359 Bremen, Germany*
[b]*Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany*
[c]*Lehrstuhl für Mikrobiologie, Technische Universität München, D-85350 Freising, Germany*
[d]*Jacobs University Bremen gGmbH, D-28759 Bremen, Germany*

Abstract

Phylogenetic analysis is currently used worldwide for taxonomic classification and identification of microorganisms. However, despite the countless trees that have been reconstructed and published in recent decades, so far, no user-friendly compilation of recommendations to standardize the data analysis and tree reconstruction process has been published. Consequently, this standard operating procedure for phylogenetic inference (SOPPI) offers a helping hand for working through the process from sampling in the field to phylogenetic tree reconstruction and publication. It is not meant to be authoritative or comprehensive, but should help to make phylogenetic inference and diversity analysis more reliable and comparable between different laboratories. It is mainly focused on using the ribosomal RNA as a universal phylogenetic marker, but the principles and recommendations can be applied to any valid marker gene. Feedback and suggestions from the scientific community are welcome in order to improve these guidelines further. Any updates will be made available on the SILVA webpage at http://www.arb-silva.de/projects/soppi.

Keywords: Phylogenetic analysis; Tree reconstruction; Ribosomal RNA; Alignment; Standardization

Introduction

Phylogenetic inference using molecular data has become a standard procedure to indentify and classify (micro)organisms. In the best case, it can even be assumed that a phylogenetic tree provides some information about the history and the ancestors of our currently existing species. It was Carl Woese, a physicist working for General Electrics, who paved the way for stable evolutionary-based systematics using comparative sequence analysis of ribosomal RNA (rRNA) molecules [18]. He propagated the idea Zuckerkandl and Pauling came up with in 1965 [19], that macromolecules, such as DNA or proteins, can be used as molecular clocks for phylogenetic inference. Thirty years have passed since Carl Woese proposed the three primary domains of life based on the phylogenetic analysis of rRNA genes. Despite ongoing discussions about the validity of the concept, the introduction of rRNA as a universal molecular marker (the "gold standard") has transformed microbiology to its roots. Cultivation-independent investigations have reported an immense array of completely unexpected microbial diversity in the environment [17]. Today, the rRNA provides the largest collection of any single molecular marker. Currently,

*Corresponding author at: Max Planck Institute for Marine Microbiology Celsiusstr. 1, D-28359 Bremen, Germany. Tel.: +49 421 2028970; fax: +49 421 2028580.
E-mail address: fog@mpi-bremen.de (F.O. Glöckner).

0723-2020/$ - see front matter © 2008 Elsevier GmbH. All rights reserved.
doi:10.1016/j.syapm.2008.08.003

more than 600,000 small (SSU) and 100,000 large (LSU) subunit rRNA variants are available in public databases, such as SILVA [14] or RDP (only SSU) [2], with the vast majority stemming from uncultured bacteria.

This immense amount of available sequence information puts pressure on the individual researcher when starting to reconstruct phylogenetic trees. Although powerful software tools for DNA sequence analysis, such as the ARB package [13], are available, the question that has arisen over and over again at workshops and conferences has been "Can't we get some guidelines to help us navigate through the sequence and tree space?" Consequently, like a protocol for laboratory experiments, this 'standard operating procedure for phylogenetic inference' (SOPPI) should help biologists to improve their workflow when working with phylogenetic marker genes. It reflects the opinion of the authors about the different steps that need to be taken into account when dealing with molecular data (mainly rRNA genes) for phylogenetic inference and tree reconstruction. However, for detailed information about the theory of phylogenetic analysis, please refer to Swofford et al. [16] and Felsenstein [7].

Sampling

Closely connected to sampling is the recording of contextual data for sequences, such as information on the sampling sites or hosts required for later data integration and interpretation [5]. Only an integrated view of our sequence collections will finally turn molecular data into biological knowledge. However, only a minority of researchers takes this into account when submitting sequence data to the public repositories. Therefore, please make sure that a minimal amount of contextual (meta)data is not only reported in the paper, but also deposited in the databases for every culture or sample devoted to sequencing of phylogenetic marker genes. For environmental samples, at least the GPS position (latitude, longitude) depth/altitude and time of sampling should be made routinely available for every (rRNA) sequence. An extended list of suggested contextual data can be found at http://www.arb-silva.de/projects/contextual-data. More information about the collaborative effort headed by the Genomics Standards Consortium to enrich the contextual data of our genome and metagenome collection, as well as environmental marker genes like rRNA, is available at http://gensc.org [9].

Sequencing

Good phylogenetic inference can only be undertaken based on high quality, nearly full-length sequences because the information content of all genetic markers is limited. For any in-depth phylogenetic rRNA analysis, a minimum length of 1200 and 1900 bases is highly recommended for 16S/18S and 23S/28S, respectively. Please avoid reporting unresolved (ambiguous) bases by rechecking the original chromatogram from the sequencing device. If this does not help, for instance, due to operon heterogeneities, unresolved bases must be handled using the standard international union of pure and applied chemistry (IUPAC) nucleotide code (see http://www.bioinformatics.org/SMS/iupac.html). Finally, do not fill up sequencing gaps with N's. Sequence quality should also be verified by taking into account the alignment that can indicate sequencing errors by considering the secondary structure and the positional variability of the molecule. Several tools exist to assist in raw sequence analysis and assembly. Examples are Sequencher (http://www.genecodes.com) and RNA-Baser (http://www.rnabaser.com). For the verification of sequence quality using secondary structure information we recommend the alignment editor of the free software package ARB (http://www.arb-home.de). Sequences between 300 and 1000 bases can be used for phylogenetic classification, but the results need to be carefully checked taking into account the reduced information content. Partial rRNA sequences shorter than 300 bases are not suited for phylogenetic inference and the only recommendation that can be given in this case is to invest additional efforts to elongate the sequences.

Alignment

Similar to high-quality sequences, a high-quality alignment is a prerequisite for any phylogenetic inference. It assures that every column represents only orthologous bases that have evolved from a common ancestor. The impact of the underlying alignment on the final tree quality should never be underestimated [12].

For rRNA alignments, it is highly recommended to take the secondary structure of the molecule into account. Orthology of the bases in the alignment can be assumed if base changes have been complemented by corresponding exchanges on both sides of the helices (Fig. 1). Experts in the field have dedicated years of their life establishing comprehensive rRNA alignments. For *Bacteria*, you can choose between alignments from SILVA (http://www.arb-silva.de, [14]) and RDP II (http://rdp.cme.msu.edu, [2]). For *Archaea* and *Eukarya*, pre-aligned sequences are only offered by the SILVA databases.

For rRNA, it is not recommended to use a subset of sequences and establish an alignment from scratch using multiple alignment tools like Clustal [1], Mafft [11] or Muscle [3].

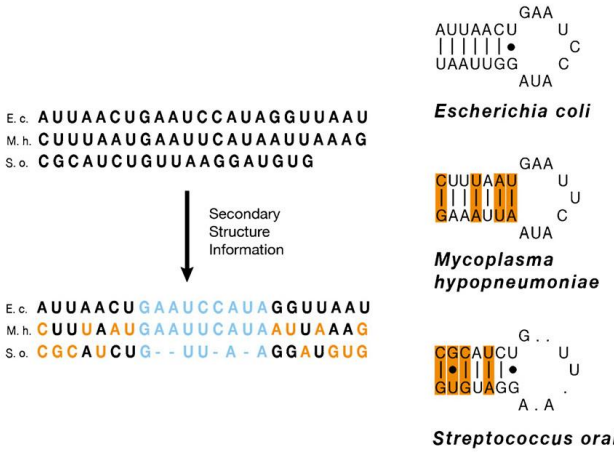

Fig. 1. The figure gives an example of how the secondary structure information of the ribosomal RNA can be used to gain confidence for positional orthology in the alignment columns. On the left side, the raw and aligned sequences for *Escherichia coli* (E. c.), *Mycoplasma hypopneumoniae* (M. h.) and *Streptococcus oralis* (S. o.) are shown. The right panel shows the respective secondary structures. The bases marked in orange indicate the corresponding base exchanges to maintain the helix structure. Dashes indicate canonical, and dots non-canonical base pairings. Especially in the case of *S. oralis*, the consideration of the pairing base pairs in the stem finally allows a reasonable alignment to be built, despite the deletions. Since the exact secondary structure of the 16S rRNA has only been determined for a few reference sequences, the *E. coli* 16S rRNA secondary structure is normally used as a model (framework) for all other bacterial sequences.

The reasons are:

1. Establishing an alignment based on the primary and secondary structure is a tedious job and it will be hard to achieve the already established standards, especially for the rRNA genes.
2. Identity values (matrices) cannot be directly compared if they are based on different alignments.
3. Resulting trees cannot be directly compared if they are based on different alignments.
4. Any alignment optimization can be conserved if you follow one of the existing alignments, for instance, the SILVA database will adopt aligned sequences for incorporation in the reference alignment (Seed) and in the next releases all related sequences will be aligned accordingly.
5. Deposition and exchange of alignments will be facilitated when adhering to existing standards.

Nevertheless, if you decide to build your own alignment, for instance, because no standard alignment is available for your current gene of interest, it is mandatory to document the alignment process exactly. This means you have to describe which method or tool (including the version of the tool) has been used and which parameters have been applied (e.g. how gaps are treated). Depending on the tool or parameters used, the resulting alignments can differ significantly [4] and this directly influences your phylogenetic or statistical analysis. Whatever you do, please take into account that the scientific community must be able to reproduce your results.

In any case, like the sequence information, the alignment needs to be made publicly available in an electronic format for evaluation and tree reproduction (ideally, this should be a multi-FASTA or ARB file including all contextual data). Please do not use cryptic headers containing your personal identifiers. Instead, use the accession number with the start and stop position to make the sequence entry unique. This is mandatory for sequences extracted from (meta)genomes. If the entry is unambiguously described by the accession number, start and stop positions can be omitted. Below is an example for an appropriate header providing the accession number, start, stop, length, and the organism name.

>AF12345678.1.1542 1542 bp *Ultrabacterium headerii*

The corresponding export filters for the ARB software suite are available at www.arb-silva.de/download/archive/imp_exp_filters.

Tree reconstruction

First of all, it is important to realize that there is no easy-to-follow standard recipe of how to produce a

publication-ready tree. The reason is, that the ultimate tree will never exist due to noise in the input data (caused by sequencing or alignment errors), limited resolution power (information content) of the genetic markers and, finally, the extremely large number of potential tree topologies (the "tree space"). Even for small subsets of sequences, not all trees can be evaluated to find the best one according to the model of evolution and tree reconstruction methods used. Testing the 2.8×10^{74} available topologies for only 50 sequences (Table 1) will take longer than the lifetime of a human being. To be able to reconstruct a tree in reasonable time, simplified evolutionary models and the application of heuristics (only a subset of tree topologies is evaluated) are accepted. Nevertheless, as a consequence, the optimal tree topology might escape from the analysis and only a suboptimal topology is reported by the algorithm.

In general, the topology of the reconstructed tree is influenced by the following factors:

1. Quality of the sequences (length, ambiguities, homopolymers)
2. Quality of the alignment (taking into account secondary structure information)
3. Amount and selection of sequences (full length, representative)
4. Amount and selection of alignment positions used for tree reconstruction (application of different filters)
5. The tree reconstruction method and parameters used.

Only if all of these parameters are taken into account and the tree topology is carefully evaluated during the analysis, a high-quality tree can finally be published.

Generally, in-depth phylogenetic analysis should only be performed with nearly full-length sequences (see above as well). The reason is that the information content of standard phylogenetic markers is already limited by its size and allowed character states per position. A further reduction of information will lead to misassignments and unstable topologies [12].

Table 1. Total number of unrooted, bifurcated trees (the "tree space") which theoretically has to be evaluated in a maximum parsimony or maximum likelihood approach to find the optimal tree, depending on the number of sequences used

Number of sequences	Number of trees
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
50	2.8×10^{74}

Calculated according to [6]. For only 50 sequences, the total number of possible trees is nearly in the range of the estimated number of atoms in our universe and trees are normally calculated using many more sequences.

Phylogenetic tree reconstruction represents nothing more than a test of the stability of the tree candidates. To do this, several tree reconstruction methods based on different algorithms need to be applied and the results compared. The three commonly used approaches are based on distance matrix (DM) (e.g. neighbour joining), maximum parsimony (MP) and maximum likelihood (ML) methods. All of these approaches have their advantages and disadvantages and a detailed explanation is beyond the scope of this SOP. To get a basic understanding of the methods and underlying evolutionary models, the book chapter "Phylogenetic Inference" by Swofford et al. [16] is recommended. At the moment, ML is regarded as the most advanced method for phylogenetic inference based on the cost of high computational demands. Nevertheless, tools like PhyML [10] or RAxML [15] provide fast ML implementations which even allow bootstrapping to be considered for ML.

Under the hypothetical assumption that the quality of the sequences and the alignment is in its best state, the amount and selection of sequences can still have a significant influence on the topology of the resulting tree(s). The typical problem that shows up in tree reconstruction is unstable branching patterns, although moving branches or species can often be stabilized by the addition of further reference sequences. This usually helps to reduce branch attraction effects caused by false identities resulting from multiple base changes during the course of evolution. Therefore, you should always go for the maximum number of sequences that can be handled within a given amount of time and computational resources. Even on a normal dual core PC, a tree with 5000 sequences can be calculated with ML in less than 2 days.

Filtering of input data (removal of selected alignment positions) is necessary to improve the signal-to-noise ratio and is an appropriate method to test the topology of the tree by altering the underlying dataset. Nevertheless, care should be taken to find the optimal threshold when removing complete alignment columns for the tree reconstruction process. Noise is superimposed on the dataset mainly by sequencing errors, false identities and a suboptimal alignment which are typically found in highly variable regions and should therefore be excluded from the calculations. However, it remains to the investigator to identify the grade of variability that should be used as a threshold for filtering, although a 50% positional conservatory filter based on the sequence identity of a specific phylogenetic group is definitely a good starting point for rRNA

genes. Such a filter will exclude the complete alignment column, if the frequency of the most abundant nucleotide is below 50%. Calculation (e.g. with the software package ARB) and testing of additional filters, such as 30% and 40% filters, are recommended in order to get a feeling for the stability of the topology. For protein genes, we recommend a 30% filter. For maximum accuracy, positional conservatory filters should always be calculated using all sequences of the corresponding phylum of your group of interest, even if your tree will finally only be calculated on a subset of sequences out of the group.

Tree reconstruction should be performed with different methods, parameters and filters, and the resulting topologies need to be compared to find out which part of the tree is stable and which one is not. This is not an easy task because no automatic method exists. A very basic possibility is to print out the different trees and mark inconsistencies to get a visual impression of the stability of the phylogenetic branchings. ARB users can also "mark" selected clusters in a tree and by switching between alternative trees it can be easily checked if the sequences under investigation show the same branching order.

Finally, how to deal with partial sequences? The worst-case scenario is to truncate full-length sequences to fit the length of the shortest sequences. If partial sequences need to be part of a tree, and extension by additional sequencing is not possible, incremental adding procedures, as offered by the ARB Parsimony system, can be used. In this case, the tree needs to be built first on full-length sequences, using MP, ML or DM, and the partial sequences are later added without allowing changes of the overall tree topology.

Tree reconstruction can be a very time consuming and computationally intensive task. If your laboratory is not prepared to deal with large datasets and different phylogenetic tools, professional assistance in publication-ready tree reconstruction and training is offered by the company Ribocon (www.ribocon.com).

Presenting the tree

For the presentation or publication of an estimated phylogeny, the tree which is considered to be the "best" (e.g. the ML tree) should be taken as the template for the introduction of multifurcations, in case, the respective branching pattern is not unambiguously supported by the different approaches [12] despite of a bifurcation. A multifurcation represents a separation from a single entity to three or even more entities. Or, in other words, the exact evolutionary history in this part of the tree remains unresolved. In the case where bootstrapping is undertaken, the respective values can be added to the unambiguously resolved branches. It is well accepted to finally show just a (small) subset of the sequences in the paper, as long as it is properly documented. This is a valid trade-off between the clarity of presentation (and the limited space of a journal page) and the accuracy of the topology, which can often only be reached with a larger dataset. To make comparisons and reproducibility of the tree easier, please avoid cryptic names on the tip of the branches. Use the accession number and the name of the organism (if available) for identification. For sequences from uncultivated organisms, try to show entries in your final tree which at least provide some additional information on the origin (e.g. "marine surface water clone" instead of "uncultivated organism"). Nevertheless, again, for your calculations you can also use high-quality "unannotated" sequences. The quality of your tree will be better the more sequences you use. In case your tree consists of sequences which significantly differ in length, the number of bases should also be provided. If you show several trees, organize them in a similar way to facilitate easy comparisons of the topologies. If you define new clusters or groups, indicate them in all trees with brackets and the corresponding cluster names.

Documentation

Irrespective of what has been carried out, the most important point is always to document accurately all steps undertaken to create the presented tree. The minimal information required to evaluate your results are the method(s), tool(s) and number of sequences that have been used for alignment and tree reconstruction, the evolutionary model (e.g. GTR, F84, JC) and the parameters (e.g. gamma distribution, filters) that have been applied and, very importantly, the number of valid columns the tree is based on (due to filtering). If multifurcations have been introduced, this needs to be clearly stated in the results and discussion, as well as in the figure legends.

Sequence submission and access to sequences

Sequences need to be submitted to one of the partners of the international nucleotide sequence database collaboration (INSDC, http://www.insdc.org), comprising DDBJ, EMBL and Genbank, to make them publicly available.

Please do not forget to add as much contextual data as possible (see paragraph on sampling) when submitting the sequences. Unfortunately, not all desired parameters can currently be deposited in the INSDC databases. Nevertheless, it is worthwhile to store them consistently, for instance, in your personal ARB/SILVA

database, and send them to the "note" field of the INSDC databases. Respective export filters for sequence and contextual data will be made available for the ARB package, therefore, please check www.arb-silva.de for forthcoming information.

Later, after acceptance of your paper, all sequences used for analysis must be immediately made publicly available via the INSDC databases. The respective accession numbers need to be clearly stated in the paper. For the review process, all sequences that are at this stage not yet available from the INSDC databases need to be sent, together with the alignments, in multi-FASTA format to the Editor of the journal considered for publication.

Example for materials and methods

"Sequences have been analysed using the ARB software package (version December 2007) [13] and the corresponding SILVA SSURef 94 database [14]. After importing, all sequences were automatically aligned according to the SILVA SSU reference alignment. Manual refinement of the alignment was carried out taking into account the secondary structure information of the rRNA. Tree reconstruction was performed with up to 1000 sequences using the neighbour joining (ARB), MP (DNAPars v1.8, [8]) and ML (RAxML v7.04, [15]) methods. Tree topology was further tested by the application of 30%, 40% and 50% positional conservatory filters. The final tree was calculated with 500 sequences based on 1280 valid columns (50% conservatory filtering) with RAxML (model: GTRGAMMA). Partial sequences have been added to the tree using the ARB parsimony tool. Multifurcations have been manually introduced in the case where tree topology could not be unambiguously resolved based on the different treeing methods and the underlying dataset. For better clarity, only selected subsets of the sequences used for treeing are shown in the figure.

The respective alignments are available in multi-FASTA format and as an ARB file at ftp.xyz.de/data. All sequences described in the paper are available from the databases of the INSDC, comprising DDBJ, EMBL and Genbank, under EX123456, EX123458 EX123470."

Acknowledgements

We thank all speakers and participants of the International Workshop on rRNA Technology, April 7 9, 2008, in Bremen, Germany for providing the final spark to compile these guidelines. This study was supported by the Max Planck Society and Ribocon GmbH.

References

[1] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T.J. Gibson, D.G. Higgins, J.D. Thompson, Multiple sequence alignment with the Clustal series of programs, Nucleic Acid Res. 31 (2003) 3497–3500.
[2] J.R. Cole, B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, A.M. Bandela, E. Cardenas, G.M. Garrity, J.M. Tiedje, The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data, Nucleic Acid Res. 35 (2007) D169–172.
[3] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 1–19.
[4] R.C. Edgar, S. Batzoglou, Multiple sequence alignment, Curr. Opin. Struct. Biol. 16 (2006) 368–373.
[5] N. Editorial, A place for everything, Nature 453 (2008), (2–2).
[6] J. Felsenstein, Number of evolutionary trees, Syst. Zool. 27 (1978) 27–33.
[7] J. Felsenstein, Inferring Phylogenies, Sinauer Associates Inc., Sunderland, MA, 2004.
[8] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.67, Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, 2005.
[9] D. Field, G. Garrity, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, T. Gray, M. Ashburner, S. Baldauf, J. Boore, G. Cochrane, J. Cole, C. dePamphilis, R. Edwards, N. Faruque, R. Feldmann, F.O. Glöckner, et al., Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification, Nat. Biotechnol. 26 (2008) 541–547.
[10] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696–704.
[11] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acid Res. 30 (2002) 3059–3066.
[12] W. Ludwig, H.P. Klenk, A phylogenetic backbone and taxonomic framework for prokaryotic systematics, in: D.R. Boone, R.W. Castenholz (Eds.), The *Archaea* and the Deeply Branching and Phototrophic *Bacteria*, Springer, New York, 2001, pp. 49–65.
[13] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A.W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, K.H. Schleifer, ARB: a software environment for sequence data, Nucleic Acid Res. 32 (2004) 1363–1371.

[14] E. Pruesse, C. Quast, K. Knittel, B.M. Fuchs, W.G. Ludwig, J. Peplies, F.O. Glöckner, SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB, Nucleic Acid Res. 35 (2007) 7188–7196.
[15] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, Bioinformatics 22 (2006) 2688–2690.
[16] D.L. Swofford, G.J. Olsen, P.J. Waddel, D.M. Hillis, Phylogenetic inference, in: D.M. Hillis, C. Moritz, B.K. Marble (Eds.), Molecular Systematics, second ed., Sinauer Associates Inc., Sunderland, MA, 1996, pp. 407–514.
[17] V. Torsvik, J. Goksoyr, F.L. Daae, High diversity in DNA of soil bacteria, Appl. Environ. Microbiol. 56 (1990) 782–787.
[18] C.R. Woese, E. Fox, Phylogenetic structure of the prokaryotic domain: the primary kingdoms, Proc. Natl. Acad. Sci. USA 74 (1977) 5088–5090.
[19] E. Zuckerkandl, L. Pauling, Molecules as documents of evolutionary history, J. Theor. Biol. 8 (1965) 357–366.

Summary

The work of this thesis resulted in a series of scholarly published articles. They can be grouped into two lines of work: a) genomic data standardization and b) software architecture development and implementation of an integrated framework for ecological genomics. The centerpiece of this thesis is the Microbial Ecological Genomics Database (MegDb) which implements the aforementioned domain model. In the vicinity of MegDb a set of tools has been developed using the ecological geo-referenced DNA sequence data.

The engineering task of this thesis has been to find a solution on how fundamental ecological questions can be formalized. Furthermore, how can data relevant to these questions be organized allowing systematic and efficient ways of analysis? In detail, several scientific questions have motivated and guided this thesis:

1. "Who is out there and where?" in terms of sequenced genomes and key genes,
2. "What are they doing?" in terms of functional capacities,
3. "Under which environmental conditions?"
4. "What is the community structure?" in terms of gene fingerprints.

These ecological questions can be posed to MegDb either directly using the Structured Query Language (SQL) or using different tools:

1. "Who is out there and where?" can be explored by using the Genes Mapserver and Geographic-Blast
2. "What are they doing?" in terms of functional capacities can be queried using SQL, MetaMine, or MetaLook
3. "Under which environmental conditions?" can be explored using the Megx.net and the Genes Mapserver
4. "What is there community structure?" in terms of gene fingerprints can be queried using SQL.

Additionally, during the course of this thesis new scientific ideas emerged leading to additional projects based on MegDb and its associated tools.

Pier Luigi Buttigieg studied the question of the existence of marine specific genes in the set of available genomes (Buttigieg 2009). Using MegDb, marine-biased distributions of genes were identified in marine genomes relative to their closest, non-marine phylogenetic

neighbors based on SEED annotations and Pfam 23.0 protein domains (Finn, Tate et al. 2008). This revealed 10 Pfam families and 80 SEED features which showed a marine-biased distribution and indicated a flexible, regulated metabolism with environmental resistance factors as characteristic of the marine bacterioplankton.

Another study carried out by Pelin Yilmaz constructed a globally representative temporally and spatially referenced 16S rDNA sequence dataset (Yilmaz 2009). This study showed that MegDb can provide accurate climatological data describing the environmental characteristics of each marine sampling location. Different oceanic locations could be grouped into distinct habitats based on a collection of abiotic parameters (temperature, salinity, dissolved oxygen concentration) and nutrient elements (nitrate, phosphate, silicate) retrieved from MegDb. This grouping of oceanic habitats provides the basis of a global survey of marine prokaryotic biogeography with respect to environmental parameters.

Overall, this thesis resulted in an integrated database suitable for ecological genomics based on existing and newly developed standards. The involvement in the Genomic Standards Consortium underpins that successful integration projects need to be based on common sense of international scientific communities.

The summary of the individual projects are detailed in the following sections.

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

At the time of its first publishing Megx.net was the first database resource offering data on marine ecological genomics (Lombardot, Kottmann et al. 2006). It was and still is a unique combination of molecular sequence data and geographic / environmental data. Geographic Information Systems (GIS) are commonly used for data integrating in the field of geology. From this point of view Megx.net represents a unique GIS on molecular sequence data. Already a plethora of genomic databases focusing on different aspects of genomic data integration exists. Megx.net is a unique integrative genomic database which for the first time includes geo-referenced sequence data and allows spatial querying. Most of the work of this thesis has been on major improvements in database structure, content and software support and standard compliance. This is summarized in the following sections.

II. Megx.net: integrated database resource for microbial ecological genomics

Most of the work of this thesis resulted in major improvements (listed below) and complete redesign of the user interface as well as the underlying system architecture of the whole system. It is planned to submit the update paper of Megx.net with the title: "Megx.net: integrated database resource for microbial ecological genomics" to the Nucleic Acids Research Database Issue 2010.

New database structure and content

Megx.net changed from a portal based on several databases, to a web accessible resource based on the single Microbial Ecological Genomics DataBase (MegDb). Megx.net follows a materialized integration approach and implements the domain model (see 17). This centralized database is managed with the PostgreSQL database management system. The geo-referenced 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)[11], and therefore offers hundreds of geospatial manipulation functions.

Currently, MegDB contains a geo-referenced collection of sequences from genomes, metagenomes, and genes of molecular environmental surveys. The initial data is retrieved from the International Nucleotide Sequence Database Collaboration (INSDC), and are further supplemented by contextual data from GOLD (Liolios, Mavromatis et al. 2008) and NCBI Genome Projects[12], and the Moore Marine Microbiology Sequencing initiative[13]. All data are manually curated. Currently, MegDb hosts draft and complete genomes, marine virus genomes, marine shotgun metagenomic datasets such as GOS, and large insert metagenomic datasets.

[11] http://www.opengeospatial.org/

[12] http://www.ncbi.nlm.nih.gov/genomes/MICROBES/ microbial_taxtree.html

[13] http://www.moore.org/microgenome/

Extension to draft genomes and shotgun datasets

The advances in sequencing technology have resulted in, an increasing number of genome and metagenome sequencing projects that are currently in progress, or stalled in a draft status (Liolios, Mavromatis et al. 2008). In January 2008, GOLD reported 3520 genome projects, of which less than one thousand were finished (Liolios, Mavromatis et al. 2008). Thus, most of the sequenced functional diversity is contained in these draft and shotgun datasets. To accommodate for this situation, megx.net was extended to host draft genomes and whole genome shotgun (WGS) datasets as well.

Further improvements:

Based on the complete redesign of Megx.net a number of features and improvements to Megx.net as published in 2006 (Lombardot, Kottmann et al. 2006) could be implemented:

- Megx.net now incorporates all sequenced marine bacteriophage genomes in MegDb
- Megx.net now includes a habitat classification for 1850 available genomes using the Habitat-Lite ontology
- MegDb is can store MIGS/MIMS contextual as recommended by the Genomics Standards Consortium (GSC)
- Megx.net provides read-only access to MIGS/MIMS reports in Genomic Contextual Data Markup Language (GCDML) XML files.
- The Genes Mapserver provides physical, chemical, geological and biological parameters, such as ocean water temperature and salinity, nutrient concentrations, organic matter and chlorophyll from World Ocean Atlas[14], World Ocean Database[15], and SeaWIFS[16] chlorophyll *a* data.
- Completely redesigned and integrated Geographic-Blast
- Access to several additional standalone tools: like MetaLook (Lombardot, Kottmann et al. 2006)., MetaMine (Bohnebeck, Lombardot et al. 2008), JCoast (Richter, Lombardot et al. 2008), TETRA(Teeling, Waldmann et al. 2004), RibAlign (Teeling and Gloeckner 2006), and MADA

[14] http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html

[15] http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html)

Some of the improvements listed above are based on individual collaborative projects which are discussed in the following chapters.

III. MetaLook: a 3D visualisation software for marine ecological genomics

MetaLook[17] offers a gene-centric view of the MegDb content, with a special focus on habitat parameters (Lombardot, Kottmann et al. 2006). It is a desktop application with a 3D user interface to interactively visualize DNA sequences on a world map. The user can define environmental containers to organize sequences according to different habitat criteria. These sequence sets can be queried by Geographic-BLAST with either genes in the database or user-imported sequences. This allows an interactive assessment of the distribution of gene functions in the environment.

The main purpose of MetaLook is to explore the possibilities of 3D visualization of high volumes of data. The choice of Java 3D as the development API, was mainly based on the decision to chose the widely installed JAVA platform. This should ease the installation and usage of MetaLook.

IV. MetaMine: A tool to detect and analyse gene patterns in their environmental context

MetaMine[18] is an interactive data mining tool which enables detection of gene patterns in an environmental context (Bohnebeck, Lombardot et al. 2008). This desktop application offers a targeted, semi-automatic search for gene patterns based on user expertise. MetaMine implements a client/server architecture to both perform BLAST searches against, and retrieve environmental data from, MegDB. The user-friendly graphical user interface allows further inspection of the calculated gene patterns in an ecological context.

Because MetaMine analyses the gene neighborhood it is mainly usable for large-insert metagenome studies, and complete and draft genomes. The short read length of 1-3 Kb in

[16] http://seawifs.gsfc.nasa.gov

[17] http://www.megx.net/metalook

[18] http://www.megx.net/metamine

Whole Genome Shotgun metagenome projects only have 1-3 genes per read which are mostly only partial. This is not adequate for the approach chosen by MetaMine. Unfortunately, the total number of nucleotide bases sequenced by Whole Genome Shotgun projects is by far higher than the number of bases from large-insert metagenome projects. Therefore, MetaMine covers only a minor set of the published sequences. However, interesting insights into the gene neighborhood can already be generated on this high quality data set and work is ongoing to evaluate more algorithms for the construction of the gene neighborhood which would also include short read fragments in the future.

V. The minimum information about a genome sequence (MIGS) specification.

An international working body of now more than 100 scientists met in a series of workshops organized by Dawn Field funded by a grant of the National Environmental Research Council (NERC) supported by the United Kingdom (Field, Garrity et al. 2005; Field, Morrison et al. 2006; Field, Garrity et al. 2007; Field, Garrity et al. 2008; Field, Glockner et al. 2008). Within the first five workshops, this group of scientists developed the recommendation of a Minimum Information about a Genome Sequence (MIGS) checklist – a minimal list of additional (contextual) data on genomic sequences (Field, Garrity et al. 2008).

The minimum information about a metagenome sequence (MIMS) extension

During the course of developing the MIGS checklist, the Max Planck Institute for Marine Microbiology joined the GSC workshops and proposed the Minimum Information about Metagenome Sequence (MIMS). Despite the addition of metagenomes as important new type of studying the genetic composition of mainly environmental samples, MIMS adds descriptors for the geographic origin and the environmental conditions present at the sampling site of the sample from which genomic material is derived and sequenced.

The initiative of the proposal was driven by the observation that existing and at that time already published metagenome studies concerned with ecological questions do not report fundamental ecological relevant information like geographic location and time of sampling. This unexpected fact renders the development of integrative systems like megx.net more complicated, because it adds the need to deal with missing data. Already handling missing data is a research topic on its own. Even if the data is given, it is usually highly dispersed in

the literature and furthermore not reported in a consistent form. Thus, just the recovering of a geographic location is a demanding process and makes integration difficult.

The MIMS addition proofed to be valuable, because it attracted the metagenomic bioinformatics platforms such as CAMERA and MG-RAST to join the GSC and lead to a joint letter to the Nature Editor (Dawn Field, Norman Morrison et al. 2008).

The rational for choosing contextual data items (descriptors throughout this text) for the MIGS specification are:

a) The descriptor of genomes and metagenomes must not yet be available or must not be mandatory for the submission of sequence data,
b) The descriptor should put the DNA sequences in a more detailed analytical research context.

Standards compliance is achieved by providing at least all the information marked as mandatory in a document of any form such as scientific articles, personal notes, excel sheets, and XML to name some among others. A reader might take the checklist and the provided document to check if all MIGS descriptors are given. This is a loose requirement on how the MIGS descriptors should be made publicly available, leaving many possible forms of reporting MIGS/MIMS compliant data.

Depending how much syntactical structure is needed, several implementation levels for creating MIGS/MIMS compliant documents exist:

1. Write down all required descriptors in a document in any form and any natural language,
2. Make a table with key/ value pairs, where the key is the name of the descriptor and value is the data provided,
3. Write down all required descriptors according to a detailed grammar for the reporting of the MIGS/MIMS data.

A clear understanding of the advantages and disadvantage of each approach helps choosing the appropriate implementation level for a given use case scenario.

The first level does not impose any constraint on how the data required by MIGS/MIMS has to be reported. It only requires that all data is somehow present in a single document. This is the lowest implementation level. It requires a human reader who reads the document and might even literally go through the MIGS/MIMS check-list and makes tick marks whenever a

MIGS/MIMS descriptor is given. This might be a reasonable level for e.g. editorial processes during article production cycles in scientific journals. With the only exception of hand written documents this level is machine readable but not machine understandable. Therefore, it would hardly be possible to automate the process of checking such a document for MIGS/MIMS compliance and would require sophisticated techniques of natural language processing.

The second level of implementation poses a simple syntactic structure on the MIGS/MIMS data in form of a list of attribute value pairs, where the attributes are the names of the descriptors and the values are the actual MIGS/MIMS data elements. This level of implementation is similar to the Entity Attribute Value (EAV) model which is discussed in detail in information integration and database communities. It is an open model because it only poses the requirement to fit everything into a key with according values (Anhoj 2003). The MIGS/MIMS check-list table appears similar to the EAV model because each row could be seen as a definition of an attribute. On the other hand, the MIGS/MIMS check-list is already a more complex model. It poses a set of rules on when which attribute needs to be applied and how. MIGS/MIMS defines rules which attributes are to be mandatory, highly desirable, and not applicable for a certain genomic study. These rules are coded in the columns of the check-list. Some additional rules are given as free text. All these rules cannot be modelled in an EAV model and would need additional effort to implement them. However, the EAV approach has the advantage to already enable a basic automation of MIGS/MIMS compliance checks by defining a list of attributes. The MIGS/MIMS rules can be checked by additional programs.

The third implementation level provides the ability to define MIGS/MIMS compliant data in very detailed, strict, ordered, and highly typed manner. Many models such as Relational Models, Object Models, Unified Model Language (UML), and XML grammars can be used for achieving the features of the third level.

From the very beginning the GSC is developing a MIGS/MIMS reference implementation using XML Schema. With the XML Schema the GSC is able to define an own mark-up language which is strongly typed and extensible. This is the most useful implementation level, because it is machine readable and processable. This reference implementation is carried out in the Genomic Contextual Data Markup Language (GCDML) project and is further discussed in the next chapter (see pp 122).

The MIGS/MIMS specification for genomes and metagenomes was published in May 2008 in Nature Biotechnology with more than 70 supporting authors.

VI. Genomic Contextual Data Markup Language (GCDML)

GCDML is a markup language for genomic contextual data defined on the basis of the Extensible Markup Language (XML). A markup language is a set of annotations to text (which can be a textual representation of data) that define how a text is to be structured, laid out, or formatted. The history of markup languages in computer science can be dated back to the 1960s when W. Tunnicliffe introduced the concept of separating information content of documents from their format. During the 1970s the Generalized Markup Language (GML) was developed. Later, work on GML resulted in the Standard Generalized Markup Language (SGML (ISO 8879:1986)) which is still widely used.

SGML is a meta-language with which specific markup languages can be defined. The most used and known language defined with SGML is the HyperText Markup Language (HTML) invented by Tim Berners Lee in 1992. HTML together with the HyperText Transfer Protocol (HTTP) lays the foundation of the modern World Wide Web[19]. The Extensible Markup Language (XML) was developed by the World Wide Web Consortium (W3C) and version 1.0 has been published in 1996. The main aim of XML is to simplify SGML by shifting the focus from professional document production by the publishing industry to documents on the Internet[20].

Therefore, XML is a simplified meta-language based on SGML that specifies syntax and rules for creating XML conforming languages.

Nowadays, XML is a widely adopted standard used in all areas of industry, government and research. Already in 2006 the XML Cover Pages listed more than 600 XML languages and it is estimated that there are more than 1000 language in use today[21]. XML also became the technology of choice for the specification of data formats for the typical office applications by the Microsoft Office Suite and Open Office. Latest, the adoption of XML by major office applications made XML ubiquitous on every desktop PC.

---

[19] http://www.w3.org/MarkUp/historical

[20] http://www.w3.org/TR/2004/REC-xml11-20040204/

[21] http://xml.coverpages.org/ and http://www.tbray.org/ongoing/When/200x/2006/01/08/No-New-XML-Languages

The wide adoption of XML is not only based on its simplicity, but also on the software support in virtually all actively developed programming languages, and many supporting standards from which the most important ones are also W3C recommendations.

Despite its wide success, XML is not the "Silver Bullet" for each and every integration or interoperability task. Several criteria should be fulfilled before the choice for XML is made. The first decision criteria concerns storage and network bandwidth needs. XML is by definition a meta-language for the markup of textual representation of data. Application scenarios could also lead to the choice of binary formats, which in principle are not human-readable, more compact and memory saving. Binary formats are often a choice where storage size and network bandwidth are critical constraints. For example, the NCBI chooses to model their data internal storage and retrieval in the Abstract Syntax Notation One (ASN.1)[22]. ASN.1 is a joint standard of the International Standard Organization and ITU-T (International Telecommunication Union – Telecommunication Standardization Sector) and is widely used in the telecommunication sector.

There has been much debate on whether ASN.1 or XML is the better technology. This resulted in attempts to develop software to bridge XML and ASN.1 for example the X.694 recommendation to map XML Schema into ASN.1[23] and the XML encoding rules (XER) to produce XML representation of data described in ASN.1[24].

However, following design constraints given in the "Architecture of the World Wide Web" by the W3C strengthen the choice of XML for the implementation of the MIGS/MIMS standard. The list of "design constraints that would suggest the use of XML include:"

1. Requirement for a hierarchical structure.
2. Need for a wide range of tools on a variety of platforms.
3. Need for data that can outlive the applications that currently process it.
4. Ability to support internationalization in a self-describing way that makes confusion over coding options unlikely.
5. Early detection of encoding errors with no requirement to "work around" such errors.
6. A high proportion of human-readable textual content.

[22] http://www.ncbi.nlm.nih.gov/Sitemap/Summary/asn1.html).

[23] http://www.itu.int/rec/T-REC-X.694/en

[24] http://www.itu.int/ITU-T/studygroups/com17/languages/X.693-0112.pdf

7. Potential composition of the data format with other XML-encoded formats.
8. Desire for data easily parsed by both humans and machines.
9. Desire for vocabularies that can be invented in a distributed manner and combined flexibly.

The GSC community demands textual representation and human readability (point 6 and 8) *a priori*. The combination of different XML languages (point 7 and 9) with the help of the Namespaces standard[25] allows defining new data formats on the basis of already existing XML language elements. Therefore, the Namespaces standard is an important technology to build up a standards stack where orthogonal standards define their own languages with their own namespaces and borrow existing elements from other languages where partial overlap exists.

In conclusion, the discussed design constraints suggest XML to be the standard of choice to define an interoperable data format for genomic contextual data which implements the MIGS/MIMS standard. Several XML schema languages for the definition of an XML language exist. The choice of XML Schema as the XML Schema language has two reasons. First, XML Schema has a rich set of built-in data types compared to all other schema languages (Murata, Lee et al. 2005) (Lee and Chu 2000). Second, XML Schema is supported by many tools. Especially because of the rich built-in types all schema binding tools support XML Schema. Schema binding tools can be used to auto-generate programming code for writing and reading GCDML files from the XML Schema itself.

Because the GSC considers supporting the use of GCDML in software applications, XML Schema is the preferred choice.

Using the Domain Model for Ecological Genomics

During the development of GCDML it turned out that the Domain Model for Ecological Genomics (see p. 17) is the most appropriate to implement the MIGS/MIMS specification. All MIGS/MIMS descriptors could be meaningful attached to one of the domain entities. Thus, arranging the descriptors in groups of study, field sample, isolate, DNA extract, clone library, DNA sequences. Although, this arrangement is not in accordance with the ordering of descriptor in the MIGS/MIMS publication, it is the only way MIGS/MIMS could be

[25] http://www.w3.org/TR/REC-xml-names/

implemented in an XML Schema so that all rules implied by MIGS/MIMS can be validated by a standard XML parser. This was not possible with the MIGS/MIMS being structured as an 'Investigation' composed of a 'Study' and an 'Assay', according to the Reporting Structures for Biological Investigations (RSBI) working group's recommendation for the modularization of checklists (Sansone, Rocca-Serra et al. 2006). Under 'Study' are the top-level concepts 'Environment' and 'Nucleic Acid Sequence' and under 'Assay' is a description of the sequencing technology. Figure 2 below depicts a refined domain model from Figure 2 (see p. 125) with all descriptors from MIGS/MIMS attached to the different domain entities.

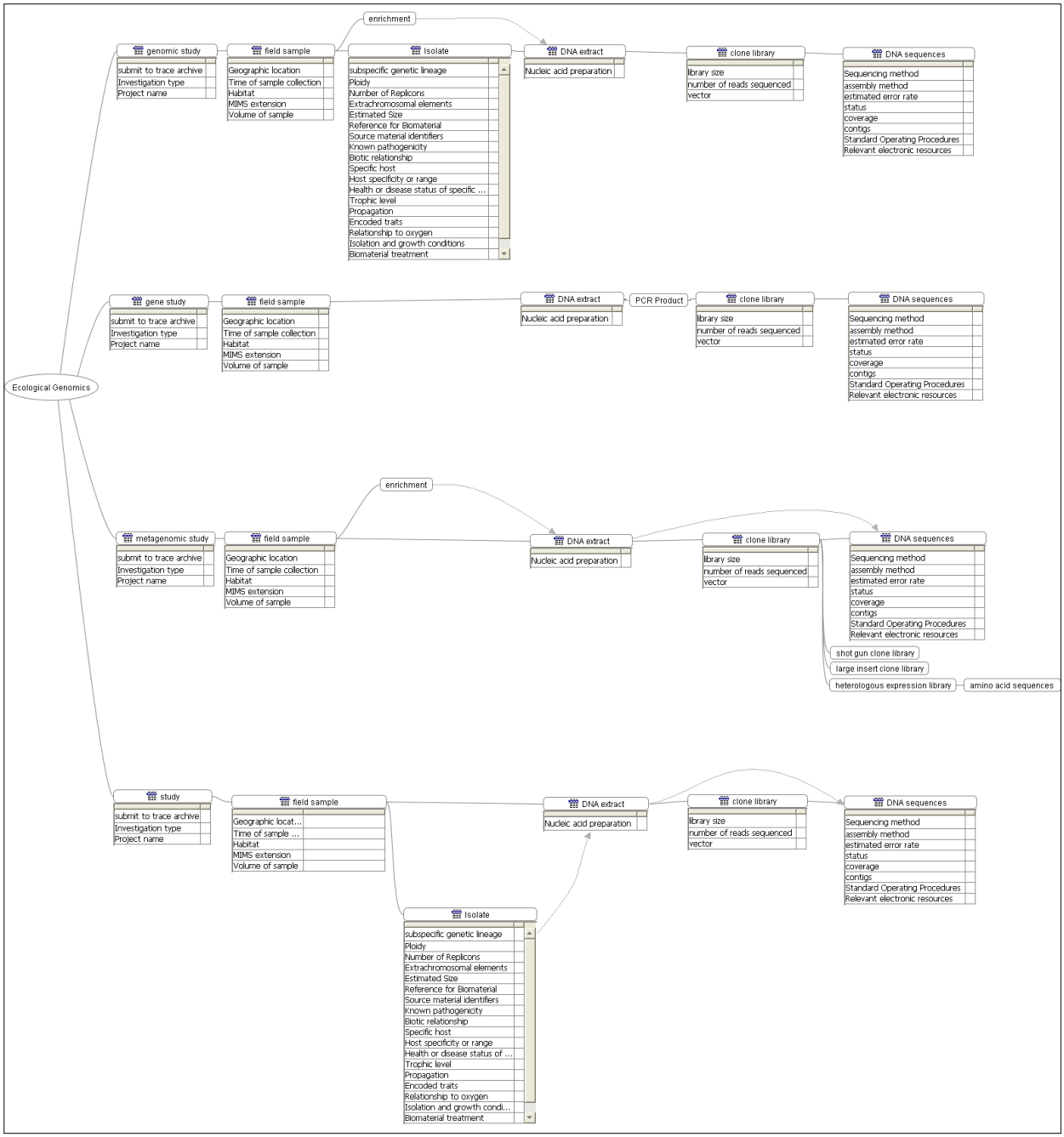

Figure 2 The Ecological Genomics Domain Model with all attributes of MIGS/MIMS added to the respective sample. Additionally a fourth line (bottom) was added to show a generalized study workflow.

Development since publication

At the time of publication GCDML version 1.5 was available. Since then, version 1.6 has been implemented and presented at the GCDML Satellite Meeting of the GSC 6 meeting in October 2008. During this two day workshop participants compiled a list of 20 bugs and feature requests. All these requests have been solved and are now included in the current 1.7 version of GCDML. Furthermore, several new main features have been implemented since the GCDML publication:

Based on community demand, GCDML introduced two clearly separated main types of reports. The MIGS report type implements only the minimal requirements of MIGS/MIMS. The "Genomic Contextual Data" (GCD) type of report extends the MIGS report type by additional descriptors not covered by the MIGS/MIMS specification and allows users to add arbitrary additional XML elements. The parallel existence of different types of reports has the advantage that the MIGS reports clearly implement nothing else but the MIGS/MIMS standard. Therefore, the schema of the MIGS reports will only change according to changes of the MIGS/MIMS specification. This allows a clearly defined and transparent synchronization between versions of the MIGS/MIMS specification and MIGS reports. On the other hand, the GCD reports allow applying changes to the extended parts not covered by the MIGS reports independently from the MIGS/MIMS specification. Thus, GCD reports give the flexibility to implement descriptors beyond common usage. As such, the GCD reports can also be used as a sandbox where new descriptors are implemented. These descriptors can later be moved to the MIGS reports if they reach common sense and be incorporated into the MIGS/MIMS specification.

The second new feature is the use of the Geography Markup Language (GML) profile only in the GCD reports. Usage of the GML profile instead of the full GML schema, allows importing only those elements from the GML namespace necessary to describe the geographic location and time of sampling in the field.

Another addition to the GCDML project is the support of automatic Java code generation from the GCDML schema using the Java Architecture for XML Binding (JAXB)[26]. An external customization file has been added to the GCDML code repository to allow users to automatically generate Java classes to read and write MIGS and GCD reports.

[26] http://jcp.org/en/jsr/detail?id=222

```
- <NasReports>
  - <MIGSReports>
    - <prokaryote gcatID="not assigned" sourceName="not assigned" sourceVersion="1">
        <ncbiOrganismName>Rhodopirellula baltica SH1</ncbiOrganismName>
        <ncbiTaxID>243090</ncbiTaxID>
        <genomeProjectID>413</genomeProjectID>
      - <originalSample>
        - <physicalMaterial>
          - <samplingTime>
              <na reason="unknown"/>
            </samplingTime>
          - <Location>
              <name>Kiel Fjord, Baltic Sea, Germany</name>
              <lat>54.329</lat>
              <lon>10.149</lon>
            </Location>
          - <amount>
              <measure uom="ml" values="100"/>
            </amount>
          - <habitat>
            - <aquatic habitatDesc="marine">
              - <waterBody>
                - <depth>
                    <measure uom="m" values="0.00 0.05"/>
                  </depth>
                </waterBody>
              </aquatic>
            </habitat>
          </physicalMaterial>
        </originalSample>
      - <isolate>
        - <isolationConditions>
            <pmid>12835416</pmid>
            <doi>10.1073/pnas.1431443100</doi>
          </isolationConditions>
        - <cultureCollection>
            <identifier>10527T</identifier>
          </cultureCollection>
          <subspecificGeneticLineage>unknown</subspecificGeneticLineage>
          <trophicLevel>heterotroph (consumer)</trophicLevel>
          <numReplicons>1</numReplicons>
          <bioticRelationship>free-living</bioticRelationship>
          <oxygenRelation>aerobic</oxygenRelation>
          <extrachromosomalElements>0</extrachromosomalElements>
        - <originalHost>
            <na reason="inapplicable"/>
          </originalHost>
          <pathogenicity>inapplicable</pathogenicity>
        </isolate>
      - <nucExtract>
          <na reason="placeholder"/>
        </nucExtract>
      - <sequencing>
          <sequencingMethod>dideoxysequencing</sequencingMethod>
          <assembly>Phrap</assembly>
          <sops/>
          <link/>
          <finishingStrategy>unknown</finishingStrategy>
        </sequencing>
      </prokaryote>
    </MIGSReports>
  </NasReports>
```

Figure 3 Example of a MIGS/MIMS compliant genome report of *Rhodopirellula baltica* (Glöckner, Kube et al. 2003) in GCDML

VII. Defining Linkages between the GSC and NSF's LTER Program: How the Ecological Metadata Language (EML) Relates to GCDML and Other Outcomes

The Ecological Metadata Language is an XML standard for data management in the Long-Term Ecological Research program (LTER), a top National Science Foundation program in the USA since 1980. Today, the LTER forms a network including over 2000 researchers associated with 26 research sites representing diverse ecosystems such as oceans, coral reefs, estuaries, lakes, deserts, prairies, alpine, and Arctic tundra, forests, urban areas, and production agriculture.

The GSC and LTER communities have independently chosen XML, XML Schema and related standards as basis for their technological developments. This gives a common technological ground, which leaves out the necessity to engineer bridges between incompatible standards and allows focusing on defining relations and finding mutual utility between GCDML and EML.

The developers of EML became interested in the MIGS/MIMS standard and GCDML, because the LTER has a growing need to adequately describe DNA sequence datasets for which no facility exists in EML yet. The EML schema is of interest to GCDML in order to avoid reinventing XML elements which are already defined by EML.

However, an important result during the discussions of this publication is a better and more detailed understanding of the different purposes and scopes EML and GCDML have.

EML has a different domain model. While the central entity for GCDML is the sample, EML models ecological metadata, which is data about ecological data (Jones, Berkley et al. 2001; Michener 2006). The metadata concept differentiates between ‘core’ data elements and additional data which give additional details to the core data elements. An often used example is a MP3 file. The binary data coding the audio signal is considered as the ‘core’ data; and the title of the encoded audio track is considered metadata which adds an additional description to the audio track.

GCDML does not make this distinction between data and metadata. The reason is to avoid confusion of what is data and what is metadata and instead emphasize the equal importance of each data element. The concept of ‘contextual data’ stresses the rational of choosing data elements which describe the research contexts within which sequences were obtained such that:

1. Comparative studies can select sequences on attributes other than raw sequence features
   e.g. Find all sequence from Baltic Sea,
2. The data generation process can be assessed:
   e.g. take only DNA sequenced with the Sanger method but not pyrosequences,
3. The data quality can be assessed
   e.g. include sequences with coverage > X.

Knowing these differences helps defining fruitful future collaborations and exchanging technological expertise and experience between the ecological and genomic communities.

VIII. Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata

This projects of the Genomic Standards Consortium aims to define a minimum workable subset of the Environment Ontology (EnvO), which is sufficient to describe the environmental origin of every DNA sequence. For the determination of this subset several text search algorithms were applied to all GenBank entries.

The list of Habitat-Lite v0.1 was used to annotate a total of 1852 organisms with a remaining 65 unclassifiable entries in MegDb. Most entries were integrated from the NCBI genome project / GOLD databases (accessed during Aug-Sep 2008). Microbes isolated from organism-associated habitats comprised 56 % of the collection while those isolated from marine waters, with a 12 % share, comprised the second largest subdivision.

This preliminary study has shown the significance of basic and easily implementable habitat classification terms applied to a large dataset. Although Habitat-Lite is a very general subset of the Environment Ontology, these terms proved sufficient to classify ~ 97 % of the 1,917 genome projects available at the onset of this project. The depth of knowledge required by the annotator is minimal in most cases, strongly reducing misclassification due to inexpert understanding of the environment / habitat terms or of the isolation procedures used in the various disciplines of microbiology. Pier Luigi Buttigieg needed less than two weeks for all Habitat-Lite assignments. This time seems to be rather short given the fact that no time was needed to learn the classification scheme and the limited set of possible classifications make the decision processes easier.

It is clear that a more refined ontology or, more realistically, a reinforcing set of orthogonal ontologies addressing isolation environment (littoral zone, epipelagic zone, coral reef, demersal zone etc.), microbial lifestyle or niche (marine snow attached, phycosphere, symbiotic, free-living etc.) and metabolic strategy (i.e. photoheterotroph, chemoautotroph, etc) would allow more apt comparative studies on a large scale. Given the compartmentalization of the genome collection by Habitat-Lite, this task becomes considerably more approachable by in-field experts and is easily implemented via MegDb.

IX. A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes

A Standard operating procedure (SOP) is a written document detailing all steps and activities of a process or procedure. The concept of SOPs originates from industry and clinical research. In clinical research, SOPs are defined by the International Conference on Harmonisation (ICH) as "detailed, written instructions to achieve uniformity of the performance of a specific function".

The publication of SOPPI is meant to give scientists producing phylogenetic trees best practice guideline at hand. Following the guidelines should help to enrich the associated data to each inferred tree and enhance reproducibility. The documentation of best practice for reporting phylogenetic inference is unique. In genomics several database and sequencing centers already adopt SOPs to increase transparency and quality of the genome annotation processes (Angiuoli, Gussman et al. 2008). In its attempt as an organization to promote standards that increase the richness and usability of genomic datasets (Field, Garrity et al. 2008), the Genomic Standards Consortium endorse the development and use of SOPs for genome annotation. Furthermore, the GSC recognizes an opportunity to create a centralized repository for SOPs (Angiuoli, Gussman et al. 2008). This would help to promote and establish better use of best practice guides within the genomics community.

Outlook

"Progress in science depends on new techniques, new discoveries and new ideas, probably in that order." – Sydney Brenner 1980

New DNA sequencing technologies have led to an enormous increase in published DNA sequences. Metagenomic analysis has enabled ecological questions to be approached on the sequence level. This has led to discoveries, such as that of the proteorhodopsin (PR) gene in 2000 suggesting the existence of non-photosynthetic phototrophy in oceans (Béjà, Aravind et al. 2000). Furthermore, the metagenomic analysis of marine surface waters conducted in the Global Ocean Sampling Expedition (GOS) confirmed the astonishing diversity of microbes while doubling the amount of sequences in the public repositories (Yooseph, Sutton et al. 2007). Further estimations on biodiversity predict that the plateau is not yet reached and still more sequence diversity can be expected in future metagenome studies (Yooseph, Sutton et al. 2007).

Current and predicted trends in the development of new sequencing technologies unanimously show that the sheer pace of sequence data growth is unlikely to slow (Hall 2007; Gupta 2008; Shendure and Ji 2008). This transforms genomics – including ecological genomics – into a data-intensive science with an exponential growth of data (Szalay and Gray 2006). Some even argue that the exponential data production is a universal fact for biology as a whole and will enable a new kind of research only limited by our capacities in computing power and bioinformatics (Szalay and Gray 2006; Committee on Metagenomics: Challenges and Functional Applications 2007). Therefore, development of data management and integration tools for any research on sequences and contextual data is a cornerstone to transfer the deluge of data into biological knowledge.

The sciences of physics and astronomy have realized their dependence on cost-effective and efficient data management and integration technologies. In "Bridging the Gap between Databases and Science" Gray and Szalay describe the successful World-Wide Telescope project (WWT) which has built a peta-byte database of astronomic data (Gray and Szalay 2004). Several factors led to the success of this project. Firstly, it was driven by 20 scientific questions which the system should be able to answer. Indeed, these 20 questions lead to the specification of the database system. This became known as the 20-questions approach and is now widely adopted in the astronomy community. Secondly, the WWT project has built a

data-centric software architecture which developed over a decade and is based on standards of the International Virtual Observatory Alliance[27]. The architecture of WWT comprises federated databases which exchange data on a regular basis.

The WWT project is a successful case study showing that it is possible to design systems capable of serving enormous amounts of data and scaling with ever more incoming data to a large scientific community.

Similar to physics and astronomy, biology needs an international community which agrees on standards as well as reliable and appropriate system architecture. The Genomic Standards Consortium gathers together representatives from the major genomic and metagenomic database and data-providers including the International Nucleotide Sequence Database Collaboration (INSDC), Genoscope, MG-RAST, CAMERA, and IMG/M among others. This consortium would be well-suited to work on a "World Wide Genomics" project to build a single architecture for the storage and analysis of the world wide, publically available DNA sequence data. This is not in the scope of the GSC at the moment. However, even without the explicit concept of a single architecture, the GSC already released several products which could be become essential components of such architecture. This consortium published the MIGS/MIMS specification and works in several other projects on more specifications of genomic data. With GCDML the GSC is implementing an XML Schema which defines an exchange format for genomic data. The more databases adopt GCDML as an exchange format the more it helps to strengthen interoperability. With GCDML and the Genomes Catalogue, the GSC is also leading software projects. The Genomes Catalogue is planned to become a central repository of MIGS/MIMS data in GCDML format and a hub for the exchange of this data. This vision places the Genomes Catalogue in a central position of a mediator in a federated system of databases especially those involved in the GSC and mentioned above.

Megx.net could play an important role in such a federated network as it is still the only resource focusing on the integration of environmental data with DNA sequences: data not covered by complementary resources such as CAMERA or MG-RAST.

Several studies using Megx.net and the underlying database have shown that the use of integrated data enables researchers to address ecologically motivated questions with *in silico*

[27] http://www.ivoa.net/Documents/

analysis of the database content. At the moment, full usage of Megx.net and MegDb can only be gained with programming and database expertise. Future work on Megx.net and MegDb has to improve the usability for non-experts. Furthermore, the scope of Megx.net should be broadened to also integrate gene function data with diversity data. This would allow researchers to query for example:

- What is the sampled diversity in terms of Operational Taxonomic Units (OTU) at a geographic location?
- What is the sampled abundance of particular OTU at a geographic location?
- Do correlations exist between and taxonomic and functional diversity at a given site?

Other future activities require data curation work. The annotation of genomes with Habitat-Lite terms allowed, for the first time, the systematic search for genomes from across a range of defined environments. Many more such annotation tasks can be done: the annotation of genomes with terms describing broad physiological capabilities like “sulphate reducer”, “sulfur oxidizer”, “nitrogen fixer” etc. would for the first time allow to query a collection of genomes with respect to physiological capacity. Another approach would be to support the annotation of environmentally relevant genes like *dsrA* or *nif* across all genomes and metagenomes to establish a reliable basis for further analysis.

Megx.net successfully integrated worldwide ocean data; however, local environmental data sources of higher resolution exist. There are several frequently sampled locations in the oceans like ALOHA and the BATS station, or Helgoland Roads. All these stations accumulate detailed environmental data and samples are routinely sequenced. Integration of the data from these stations would give a higher resolution of environmental data in space and time. Ongoing projects like MIMAS seek to utilize these high resolution data to study the microbial community structure over time. This combination is very likely to provide insights into the microbial community function in response to environmental gradients.

With Megx.net, MegDb, MetaLook, MetaMine, MIGS/MIMS, GCDML, Habitat-Lite, SOPPI and the development of other software and standards in this work, new techniques to better make sense of the data deluge are available. Based on this, new discoveries can be made and a set of new ideas for further enhancements already exist. In the course of this work it became clear that progress heavily depends on biocuration to guaranty high data quality, data integration and expert human resources.

Publication List

List of publication with contributions not further discussed in this thesis.

JCoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes.

Authors: Richter, Michael, Thierry Lombardot, Ivaylo Kostadinov, Renzo Kottmann, Melissa Beth Duhaime, Jörg Peplies, and Frank Oliver Glöckner.

Published in *BMC Bioinformatics* 9: 177.

Cataloguing Our Complete Genome Collection III.

Authors: Dawn Field, George Garrity, Tanya Gray, Jeremy Selengut, Peter Sterk, Nick Thomson, Tatiana Tatusova, Guy Cochrane, Frank Oliver Glöckner, Renzo Kottmann *et al*.

Published in *Comparative and Functional Genomics*. 2007;7.

Meeting Report: The Fourth Genomic Standards Consortium (GSC) Workshop.

Authors: Dawn Field, Frank Oliver Glöckner, George M. Garrity, Tanya Gray, Peter Sterk, Guy Cochrane, Robert Vaughan, Eugene Kolker, Renzo Kottmann *et al.*

Pulished in *OMICS*. 2008; 12(2): 101-108.

Meeting Report: The Fifth Genomic Standards Consortium (GSC) Workshop.

Authors: Dawn Field, George M. Garrity, Susanna-Assunta Sansone, Peter Sterk, Tanya Gray, Nikos Kyrpides, Lynette Hirschman, Frank Oliver Glöckner, Renzo Kottmann *et al.*

Published in *Omics*. 2008; 12(2): 109-113.

Working together to put molecules on the map.

Authors: Dawn Field, Norman Morrison, Frank Oliver Glöckner, Renzo Kottmann *et al.*

Published in *Nature*. 2008; 453(7198): 978.

Acknowledgements

Foremost, I would like to acknowledge Prof. Dr. Frank Oliver Glöckner for accepting me as a PhD student and his great and friendly support during this exciting time. The European Union and the Max Planck Society are acknowledged for funding this thesis. Thanks go to all members of the Metafunctions team, especially to Johanna Wesnigk for the great collaboration. I would also like to thank the International Max Planck Research School for Marine Microbiology especially Prof. Dr. Rudolf Amann for accepting me "as an experiment" and giving me the opportunity to study in a highly interdisciplinary environment.

I am very grateful to Dr. Thierry Lombardot for his supervision in the early years of my thesis and teaching me to see technology development as a means of solving biological questions. Thanks Dr. Michael Richter for his all-day support and fruitful discussions.

Thanks are due to the megx.net team: Ivalyo Kostadinov, Melissa Beth Duhaime, Frank Oliver Glöckner, Wolfgang Hankeln, Pelin Yilmaz, and Pier Luigi Buttigieg. The order of names does not imply anything. Thanks go to all members equally, because all of you show how productive and inspiring team work can be for all us. Nevertheless, Melissa and Ivo, we share an office. Both of you know, why I thank each of you especially.

I would like to acknowledge all members of the Microbial Genomics Group for many interesting discussions and practical support.

The Molecular Ecology group is thanked for keeping me in touch with the laboratory work and way of thinking.

I would like to thank all members of the Genomic Standards Consortium for great and productive discussions. Special thanks to Dawn Field, Saul Kravitz, Tanya Gray, and Peter Sterk.

Aleksandar Pop Ristov is thanked for introducing me to the world of business and large software development. Especially, I would like thank Mr. Pop Ristov for letting some of his best programmers develop software for this thesis for free. This showed me how industrial software production greatly improves scientific software.

I am deeply indebted to my Mother and Frank for their love and strong belief in me.

Petra, I will not only mention your name, but take the chance to express my greatest thanks. Boo, te sakam, shmok.

References

Agency, E. S. (2004). "ESA - Space Science - How many stars are there in the Universe?" Retrieved 25.04.2009, 2009, from http://www.esa.int/esaSC/SEM75BS1VED_index_0.html.

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

Angiuoli, S. V., A. Gussman, et al. (2008). "Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation." OMICS **12**(2): 137-41.

Anhoj, J. (2003). "Generic design of Web-based clinical databases." J Med Internet Res **5**(4): e27.

Baxevanis, A. D. (2008). "Searching NCBI databases using Entrez." Curr Protoc Bioinformatics **Chapter 1**: Unit 1 3.

Béjà, O., L. Aravind, et al. (2000). "Bacterial rhodopsin: evidence for a new type of phototrophy in the sea." Science **289**(5486): 1902—1906.

Benson, D. A., I. Karsch-Mizrachi, et al. (2009). "GenBank." Nucleic Acids Res **37**(Database issue): D26-31.

Berman, H. M. (2008). "The Protein Data Bank: a historical perspective." Acta Crystallogr A **64**(Pt 1): 88-95.

Bohnebeck, U., T. Lombardot, et al. (2008). "MetaMine - A tool to detect and analyse gene patterns in their environmental context." BMC Bioinformatics **9**(1): 459.

Bourne, P. E., J. Westbrook, et al. (2004). "The Protein Data Bank and lessons in data management." Brief Bioinform **5**(1): 23-30.

Bult, C. J., O. White, et al. (1996). "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii." Science (New York, N.Y.) **273**(5278): 1058-1073.

Buttigieg, P. L. (2009). Investigating habitat-specific signatures of the marine bacterioplankton: insights from comparative genomics. Max Planck Institute for Marine Microbiology. Bremen. **MSc**.

Cochrane, G., R. Akhtar, et al. (2008). "Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database." Nucleic Acids Res **36**(Database issue): D5-12.

Cole, J. R., B. Chai, et al. (2007). "The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data." Nucleic Acids Res **35**(Database issue): D169—D172.

Committee on Metagenomics: Challenges and Functional Applications, N. R. C. (2007). The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet, The National Acadamies Press, Washington DC.

Consortium, U. (2009). "The Universal Protein Resource (UniProt) 2009." Nucleic Acids Res **37**(Database issue): D169-74.

Curtis, T. P., W. T. Sloan, et al. (2002). "From the Cover: Estimating prokaryotic diversity and its limits." PNAS **99**(16): 10494-10499.

Dawn Field, Norman Morrison, et al. (2008). "Working together to put molecules on the map." Nature **453**(7198): 978.

del Giorgio, P. A. and C. M. Duarte (2002). "Respiration in the open ocean." Nature **420**(6914): 379-384.

DeLong, E. F. (1992). "Archaea in Coastal Marine Environments." PNAS **89**(12): 5685—5689.

DeLong, E. F. (2005). "Microbial community genomics in the ocean." Nat Rev Micro **3**(6): 459-469.

DeSantis, T. Z., P. Hugenholtz, et al. (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Appl Environ Microbiol **72**(7): 5069—5072.

Etzold, T., A. Ulyanov, et al. (1996). "SRS: information retrieval system for molecular biology data banks." Methods Enzymol **266**: 114-28.

Field, C. B., M. J. Behrenfeld, et al. (1998). "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components." Science **281**(5374): 237-240.

Field, D., G. Garrity, et al. (2008). "The minimum information about a genome sequence (MIGS) specification." Nat Biotech **26**(5): 541-547.

Field, D., G. Garrity, et al. (2007). "eGenomics: Cataloguing Our Complete Genome Collection III." Comparative and Functional Genomics **2007**: 7.

Field, D., G. Garrity, et al. (2005). "eGenomics: Cataloguing our Complete Genome Collection." Comp Funct Genomics **6**(7-8): 363-8.

Field, D., G. M. Garrity, et al. (2008). "Meeting Report: The Fifth Genomic Standards Consortium (GSC) Workshop." Omics **12**(2): 109-113.

Field, D., F. O. Glockner, et al. (2008). "Meeting Report: The Fourth Genomic Standards Consortium (GSC) Workshop." Omics **12**(2): 101-108.

Field, D., N. Morrison, et al. (2006). "Meeting report: eGenomics: Cataloguing our Complete Genome Collection II." OMICS **10**(2): 100-4.

Finn, R. D., J. Tate, et al. (2008). "The Pfam protein families database." Nucl. Acids Res. **36**(suppl_1): D281-288.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Fuhrman, J. A., K. McCallum, et al. (1993). "Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans." Appl Environ Microbiol **59**(5): 1294-302.

Garcia-Molina, H., J. D. Ullman, et al. (2002). Database Systems: The Complete Book. New Jersey, Prentice Hall.

Gevers, D., F. M. Cohan, et al. (2005). "Re-evaluating prokaryotic species." Nat Rev Micro **3**(9): 733-739.

Giovannoni, S. J., T. B. Britschgi, et al. (1990). "Genetic diversity in Sargasso Sea bacterioplankton." Nature **345**(6270): 60—63.

Glöckner, F. O., M. Kube, et al. (2003). "Complete genome sequence of the marine

planctomycete Pirellula sp. strain 1." Proceedings of the National Academy of Sciences of the United States of America **100**(14): 8298-8303.

Glöckner, F. O. and A. Meyerdierks, Eds. (2005). Metagenome Analysis.

Goble, C. and R. Stevens (2008). "State of the nation in data integration for bioinformatics." Journal of Biomedical Informatics **41**(5): 687-693.

Gonzalez, J. M. and M. A. Moran (1997). "Numerical dominance of a group of marine bacteria in the alpha-subclass of the class Proteobacteria in coastal seawater." Appl Environ Microbiol **63**(11): 4237-42.

Gray, J. and A. Szalay (2004). "Where the Rubber Meets the Sky: Bridging the Gap between Databases and Science." IEEE Data Engineering Bulletin **27**(4): 8.

Gupta, P. K. (2008). "Single-molecule DNA sequencing technologies for future genomics research." Trends Biotechnol **26**(11): 602-11.

Hall, N. (2007). "Advanced sequencing technologies and their wider impact in microbiology." J Exp Biol **210**(Pt 9): 1518-25.

Hernandez, T. and S. Kambhampati (2004). "Integration of biological sources: current systems and challenges ahead." SIGMOD Rec. **33**(3): 51-60.

Hugenholtz, P. (2002). "Exploring prokaryotic diversity in the genomic era." Genome Biology **3**(2): reviews0003.1-reviews0003.8.

Hugenholtz, P., B. M. Goebel, et al. (1998). "Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity." J. Bacteriol. **180**(24): 6793.

INSDC. (2009). "Feature Table Definition." Retrieved 25.04.2009, from http://www.insdc.org/files/documents/feature_table.html.

Jones, M. B., C. Berkley, et al. (2001). "Managing scientific metadata." Internet Computing, IEEE **5**: 59-68.

Jorgensen, B. B. (1982). "Mineralization of organic matter in the sea bed[mdash]the role of sulphate reduction." Nature **296**(5858): 643-645.

Juncker, A., L. Jensen, et al. (2009). "Sequence-based feature prediction and annotation of proteins." Genome Biology **10**(2): 206.

Keller, V. E. F. and L. L. Winship (2002). The Century of the Gene, Harvard Univ Pr.

Kulikova, T., R. Akhtar, et al. (2007). "EMBL Nucleotide Sequence Database in 2006." Nucleic Acids Res **35**(Database issue): D16-20.

Kyrpides, N. "GOLD - Genomes OnLine Database." from http://www.genomesonline.org/index.htm.

Lee, D. and W. W. Chu (2000). "Comparative Analysis of Six XML Schema Languages." ACM SIGMOD RECORD **29**: 76--87.

Liolios, K., K. Mavromatis, et al. (2008). "The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata." Nucl. Acids Res. **36**(suppl_1): D475-479.

Lombardot, T., R. Kottmann, et al. (2006). "Megx.net—database resources for marine ecological genomics." Nucl. Acids Res. **34**(suppl_1): D390-393.

Ludwig, W., O. Strunk, et al. (2004). "ARB: a software environment for sequence data." Nucl. Acids. Res. **32**(4): 1363-1371.

Madigan, M. T., J. Martinko, et al. (2002). Brock Biology of Microorganisms, Prentice Hall.

Markowitz, V. M. (2007). "Microbial genome data resources." Curr Opin Biotechnol **18**(3): 267—272.

Markowitz, V. M., N. N. Ivanova, et al. (2008). "IMG/M: a data management and analysis system for metagenomes." Nucl. Acids Res. **36**(suppl_1): D534-538.

Markowitz, V. M., E. Szeto, et al. (2008). "The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions." Nucleic Acids Res **36**(Database issue): D528—D533.

Médigue, C. and I. Moszer (2007). "Annotation, comparison and databases for hundreds of bacterial genomes." Research in Microbiology **158**(10): 724-736.

Meyer, F., D. Paarmann, et al. (2008). "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes." BMC Bioinformatics **9**: 386.

Michener, W. K. (2006). "Meta-information concepts for ecological data management." Ecological Informatics **1**(1): 3-7.

Murata, M., D. Lee, et al. (2005). "Taxonomy of XML schema languages using formal language theory." ACM Trans. Interet Technol. **5**(4): 660-704.

Nyren, P. (2007). "The history of pyrosequencing." Methods Mol Biol **373**: 1-14.

Olsen, G. J., D. J. Lane, et al. (1986). "Microbial ecology and evolution: a ribosomal RNA approach." Annu Rev Microbiol **40**: 337-65.

Pedrós-Alió, C. (2006). "Genomics and marine microbial ecology." International Microbiology: The Official Journal of the Spanish Society for Microbiology **9**(3): 191-7.

Pomeroy, L. R. (1974). "The Ocean's Food Web, A Changing Paradigm." BioScience **24**(9): 499-504.

Pruesse, E., C. Quast, et al. (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucleic Acids Research **35**(21): 7188-96.

Reeburgh, W. S. (1996). "Soft Spots" in the global methane budget. Microbial Growth on C1 Compounds. M. E. Lidstrom and F. R. Tabita, Kluwer**:** 334-342.

Rentzsch, R. and C. A. Orengo (2009). "Protein function prediction – the power of multiplicity." Trends in Biotechnology **27**(4): 210-219.

Richter, M., T. Lombardot, et al. (2008). "JCoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes." BMC Bioinformatics **9**: 177.

Riesenfeld, C. S., P. D. Schloss, et al. (2004). "METAGENOMICS: Genomic Analysis of Microbial Communities." Annual Review of Genetics **38**(1): 525-552.

Ronaghi, M. (2001). "Pyrosequencing sheds light on DNA sequencing." Genome Res **11**(1): 3-11.

Rosello-Mora, R. and R. Amann (2001). "The species concept for prokatyotes." FEMS Microbiol Rev. **25**(1): 39-67.

Saiki, R. K., D. H. Gelfand, et al. (1988). "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase." Science **239**(4839): 487-91.

Sansone, S.-A., P. Rocca-Serra, et al. (2006). "A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group." OMICS **10**(2): 164—171.

Sayers, E. W., T. Barrett, et al. (2009). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **37**(Database issue): D5-15.

Scheibye-Alsing, K., S. Hoffmann, et al. (2009). "Sequence assembly." Computational Biology and Chemistry **33**(2): 121-136.

Schmidt, T. M., E. F. DeLong, et al. (1991). "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing." J Bacteriol **173**(14): 4371-8.

Seshadri, R., S. A. Kravitz, et al. (2007). "CAMERA: a community resource for metagenomics." PLoS Biol **5**(3): e75.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-45.

Sterk, P., T. Kulikova, et al. (2007). "The EMBL Nucleotide Sequence and Genome Reviews Databases." Methods Mol Biol **406**: 1-21.

Straalen, V. N. M. v. and D. Roelofs (2006). An introduction to ecological genomics, Oxford University Press.

Streit, W. R. and R. A. Schmitz (2004). "Metagenomics - the key to the uncultured microbes." Current Opinion in Microbiology **7**: 492-498.

Sugawara, H., K. Ikeo, et al. (2009). "DDBJ dealing with mass data produced by the second generation sequencer." Nucleic Acids Res **37**(Database issue): D16-8.

Suzuki, M. T., M. S. Rappe, et al. (1997). "Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample." Appl. Environ. Microbiol. **63**(3): 983-989.

Szalay, A. and J. Gray (2006). "2020 Computing: Science in an exponential world." Nature **440**(7083): 413-414.

Teeling, H. and F. O. Gloeckner (2006). "RibAlign: a software tool and database for eubacterial phylogeny based on concatenated ribosomal protein subunits." BMC Bioinformatics **7**: 66.

Teeling, H., J. Waldmann, et al. (2004). "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences." BMC Bioinformatics **5**(1): 163.

Torsvik, V., J. Goksoyr, et al. (1990). "High diversity in DNA of soil bacteria." Appl Environ Microbiol **56**(3): 782-7.

Turnbaugh, P. J., M. Hamady, et al. (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-4.

Tyson, G. W. and J. F. Banfield (2005). "Cultivating the uncultivated: a community genomics perspective." Trends Microbiol **13**(9): 411-5.

Van Brabant, B., T. Gray, et al. (2008). "Laying the foundation for a Genomic Rosetta Stone: creating information hubs through the use of consensus identifiers." OMICS **12**(2): 123-7.

Vosberg, H. P. (1989). "The polymerase chain reaction: an improved method for the analysis of nucleic acids." Hum Genet **83**(1): 1-15.

Wheeler, D. L., D. M. Church, et al. (2004). "Database resources of the National Center for Biotechnology Information: update." Nucl. Acids Res. **32**(90001): D35-40.

Whitman, W. B., D. C. Coleman, et al. (1998). "Prokaryotes: the unseen majority." Proc Natl Acad Sci U S A **95**(12): 6578-83.

Widdel, F. and T. A. Hansen (1992). The Prokaryotes: A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Springer Verlag, New York. **1**.

Wiederhold, G. (1992). "Mediators in the architecture of future information systems." Computer **25**(3): 38-49.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." Proceedings of the National Academy of Sciences of the United States of America **74**(11): 5088-5090.

Yilmaz, P. (2009). Application of statistical methods for the analysis of geo-referenced sequence data in their environmental context. Max Planck Institute for Marine Microbiology. Bremen. **Msc**.

Yooseph, S., G. Sutton, et al. (2007). "The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families." PLoS Biol **5**(3): e16.

Zdobnov, E. M., R. Lopez, et al. (2002). "The EBI SRS server-new features." Bioinformatics **18**(8): 1149-50.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." Journal of Theoretical Biology **8**(2): 357-366.