UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1388

Molecular methods for microbial ecology

Developments, applications and results

LUCAS SINCLAIR

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2016

ISSN 1651-6214
ISBN 978-91-554-9620-3
urn:nbn:se:uu:diva-297613

Dissertation presented at Uppsala University to be publicly examined in Friessalen, Evolutionary Biology Center, Norbyvägen 14, Uppsala, Thursday, 8 September 2016 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Associate Professor Josh Neufeld (Waterloo University).

Abstract

Sinclair, L. 2016. Molecular methods for microbial ecology. Developments, applications and results. (Systems approach to functional characterization of lentic systems). *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1388. 52 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9620-3.

Recent developments in DNA sequencing technology allow the study of microbial ecology at unmatched detail. To fully embrace this revolution, an important avenue of research is the development of bioinformatic tools that enable scientists to leverage and manipulate the exceedingly large amounts of data produced. In this thesis, several bioinformatic tools were developed in order to process and analyze metagenomic sequence data. Subsequently, the tools were applied to the study of microbial biogeography and microbial systems biology.

A targeted metagenomics pipeline automating quality filtering, joining and taxonomic annotation was developed to assess the diversity of bacteria, archaea and eukaryotes permitting the study of biogeographic patterns in great detail. Next, a second software package which provides annotation based on environmental ontology terms was coded aiming to exploit the cornucopia of information available in public databases. It was applied to resource tracking, paleontology, and biogeography. Indeed, both these tools have already found broad applications in extending our understanding of microbial diversity in inland waters and have contributed to the development of conceptual frameworks for microbial biogeography in lotic systems. The programs were used for analyzing samples from several environments such as alkaline soda lakes and ancient sediment cores. These studies corroborated the view that the dispersal limitations of microbes are more or less non-existant as environmental properties dictating their distribution and that dormant microbes allow the reconstruction of the origin and history of the sampled community.

Furthermore, a shotgun metagenomics analysis pipeline for the characterization of total DNA extraction from the environment was put in place. The pipeline included all essential steps from raw sequence processing to functional annotation and reconstruction of prokaryotic genomes. By applying this tool, we were able to reconstruct the biochemical processes in a selection of systems representative of the tens of millions of lakes and ponds of the boreal landscape. This revealed the genomic content of abundant and so far undescribed prokaryotes harboring important functions in these ecosystems. We could show the presence of organisms with the capacity for photoferrotrophy and anaerobic methanotrophy encoded in their genomes, traits not previously detected in these systems. In another study, we showed that microbes respond to alkaline conditions by adjusting their energy acquisition and carbon fixation strategies. To conclude, we demonstrated that the "reverse ecology" approach in which the role of microbes in elemental cycles is assessed by genomic tools is very powerful as we can identify novel pathways and obtain the partitioning of metabolic processes in natural environments.

Keywords: bioinformatics, microbiology, metagenomics, ecology, metabolism

Lucas Sinclair, , Department of Ecology and Genetics, Limnology, Norbyv 18 D, Uppsala University, SE-75236 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-9620-3
urn:nbn:se:uu:diva-297613 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-297613)

To my uncle,
Thanks Frankie!

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Lucas Sinclair, Omneya Osman, Stephan Bertilsson, Alexander Eiler. (2015) **Microbial community composition and diversity: evaluating the Illumina platform.** *Published in PLOS ONE (doi:10.1371/journal.pone.0116955).*

II Lucas Sinclair, Umer Ijaz, Lars Juhl Jensen, Evangelos Pafilis, Christopher Quince, et al. (2016) **Seqenv: linking microbes to environments through text mining.** *Submitted to PeerJ (doi:10.7287/peerj.preprints.2317v1).*

III Domenico Savio, Lucas Sinclair, Umer Ijaz, Alexander Kirschner, Andreas Farnleitner, Alexander Eiler, et al. (2015) **Bacterial diversity along a 2'600 km river continuum.** *Published in Environmental Microbiology (doi:10.1111/1462-2920.12886).*

IV Lucas Sinclair, Sari Peura, Moritz Buck, Pilar Hernández, Martha Schattenhofer, Alexander Eiler. (2016) **Photoferrotrophy and anaerobic methanotrophy under-ice in boreal lakes.** *Submitted to ISME Journal.*

V Lucas Sinclair, Sainur Samad, Alexander Kirschner, Alexander Eiler. (2016) **Ecogenomics of microbes along pH gradients in alkaline soda lakes.** *Manuscript.*

List of additional papers

Outside this thesis, the author has contributed to the following work.

- Sari Peura, Lucas Sinclair, Stefan Bertilsson, Alexander Eiler. (2015) **Metagenomic insights into strategies of aerobic and anaerobic carbon and nitrogen transformation in boreal lakes.** *Published in Scientific Reports (doi:10.1038/srep12102).*

- Alexander Eiler, Rhiannon Mondav, Lucas Sinclair, Leyden Vidal, Douglas Scofield, Patrick Schwientek et al. (2016) **Tuning fresh: radiation through rewiring of central metabolism in streamlined bacteria.** *Published in ISME Journal (doi:10.1038/ismej.2015.260).*

- Silke Langenheder, Jérôme Comte, Yinghua Zha, Sainur Samad, Lucas Sinclair, Alexander Eiler, Eva Lindström. (2016) **Remnants of marine bacterial communities can be retrieved from deep sediments in lakes of marine origin.** *Published in Environmental Microbiology Reports (doi:10.1111/1758-2229.12392).*

- Sari Peura, Moritz Buck, Lucas Sinclair, Sanni Aalto, Hannu Nykänen, Alexander Eiler. (2016) **Autotrophy along a natural redox tower: Novel phototrophs and chemotrophs in a boreal lake.** *Submitted to Microbiome.*

- Kemal Sanli, Lucas Sinclair, Henrik Nilsson, Adil Mardinoglu, Alexander Eiler. (2016) **PACFM: Pathway Analysis with Circos for Functional Metagenomics.** *Submitted to BMC Bioinformatics.*

Contents

1 Prologue 13
2 Introduction 15
3 Aims 30
4 Methods 32
5 Results and discussion 38
6 Conclusion and perspectives 45
7 Summary in Swedish 48
8 Acknowledgements 51

Paper I 57

Paper II 85

Paper III 107

Paper IV 133

Paper V 161

References 178

List of Figures

Figure 1: Schematic representation of the great plate count anomaly. 16
Figure 2: Cost for sequencing a million nucleotides over time. 18
Figure 3: Discrete or continuous units of biological classification. 21
Figure 4: Viewing a FASTA formated sequence file in a plain text editor. 33
Figure 5: Different methods for microbial ecology. Adapted from [1]. 47
Figure 6: Distribution of barcodes matching and mismatching. 72
Figure 7: Distribution of sequence lengths for matching barcodes. 73
Figure 8: Composition of different lengths fractions. 74
Figure 9: Correspondence of phylogenetic distance methods. 75
Figure 10: Experiment outline. 77
Figure 11: Pipeline outline. 78
Figure 12: Taxa relative abundance comparing 454 and Illumina samples. 80
Figure 13: NMDS comparing 454 and Illumina samples. 81
Figure 14: Schematic of the internal functioning of the seqenv pipeline. 96
Figure 15: The EnvO terms associated with two AOA OTUs. 98
Figure 16: EnvO terms and OTU richness against mean OTU pH. 99
Figure 17: Heatmap of top ten EnvO terms for determining OTU pH. 100
Figure 18: NMDS plot of Black Sea plankton 18S rRNA samples. 101
Figure 19: Heatmap of top ten EnvO terms for determining ES. 102
Figure 20: Danube geography and sampling sites. 121
Figure 21: Core communities in the Danube. 122
Figure 22: Danube bacterial richness and evenness. 123
Figure 23: Danube geography and sampling sites. 124
Figure 24: Dynamics of most abundant freshwater tribes. 125
Figure 25: Results from the seqenv analysis. 126
Figure 26: Relative abundance of the main phyla. 130
Figure 27: Summary of lake characteristics. 148
Figure 28: Bacterial community profiles by lake and by depth. 149
Figure 29: Genome encoded metabolic profiles by depth. 150
Figure 30: Traits, taxonomy and phylogeny of reconstructed genomes. 151
Figure 31: Main degradation pathways of allochthonous organic matter. 152
Figure 32: Cell counts per depth for all lakes. 155
Figure 33: Chemistry and gas profiles in lakes RL and SB. 156
Figure 34: Ten most abundant OTUs for lakes RL and SB. 157
Figure 35: NMDS based on PFAMs found in three lakes. 158
Figure 36: Map of the different soda lakes around Neusiedler. 173
Figure 37: Environmental and microbial dynamics. 174
Figure 38: Traits and phylogeny of reconstructed genome bins. 175

Abbreviations

16S	16 Svedberg sedimentation mark (non-SI unit)
AA	Amino acid
ACE	Abundance-based coverage estimators (diversity)
ANOVA	Analysis of variance
AOA	Ammonia oxidising archaea
BLAST	Basic local alignment search tool
BP	Base pair (of nucleotides)
DIC	Dissolved inorganic carbon
DNA	Deoxyribonucleic acid
DOC	Dissolved organic carbon
EnvO	Environmental ontology
GHG	Green house gas
GI	GenInfo identifier
HMM	Hidden Markov model
HPLC	High performance liquid chromatography
HTS	High-throughput (genetic) sequencing
IT	Information technology
k-mer	All the possible substrings of length k (in a string)
MAG	Metagenome assembled genome
MIDs	Multiplex identifiers
mRNA	Messenger ribonucleic acid
MSA	Multiple sequence alignement
MSE	Mean squared error
NADH	Nicotinamide adenine dinucleotide (reduced)
NCBI	National center for biotechnology information (in the US)
NER	Named entity recognition
NMDS	Non-parametric multidimensional scaling (ordination plot)
OTU	Operational Taxonomic Unit
PCR	Polymerase chain reaction
PDF	Portable document format
PFAM	A protein family represented by a MSA and a HMM
RAM	Random access memory
RKM	River kilometer (of the Danube)
rRNA	Ribosomal ribonucleic acid
SMF	Sodium-motive force
TOC	Total organic carbon
TON	Total organic nitrogen
TSV	Tab separated values

1. Prologue

This thesis revolves around the development of bioinformatic methods and tools for the subsequent application to the field of microbial ecology. The science of microbial ecology aims to understand and predict the relationships between microscopic organisms as well as their interactions with the surrounding environment. Of course, given the small size and inability to manipulate the objects of study, there are tough methodological constraints when one attempts to carry out microbial ecology. For example, microbiologists have remained largely unsuccessful, over the last century, in isolating and growing the majority of species that are found in the wild. Indeed, less than 0.1% of microbes multiply and form colonies in a petri dish on a nutrient-loaded agar plate, no matter how much we tune the artificial conditions [2].

Nowadays, to overcome some of these limitations, molecular methods have become widely applied, and, of particular popularity, is the approach of high-throughput genetic sequencing (HTS). These expensive laboratory tools, when given unsorted solutions of DNA fragments and the correct reagents as input, are able to <read> the genetic contents written in the universal genetic alphabet composed of the four letters, A, T, G and C. In turn, they output large quantities of digital information in the form of computer files.

Total environmental DNA sequencing bypasses many of the obstacles of classical microbiology and has transformed microbial ecology into an information science over the last decades, in a smiliar fashion as the field of particle physics which was forced into a data-driven *modus operandi*. Rapidly, algorithms and automation through programming have become a large and mandatory part of the analysis workflow, requiring new skill sets from scientists. Without the construction of software pipelines to produce the interpretable visualizations and compute the hypothesis-testing statistics, the endeavor would be vain.

To understand the type of questions that these tools can help us answer, we need to briefly explain what they are able to measure in more detail. There are two main types of HTS strategies in microbial ecology:

The first is to target a specific gene that is essential for vital functions and is thus present in all living microorganisms. This approach is aptly named targeted-metagenomics and, by restricting the sequencing to only one region of the genomes, enables us to qualitatively measure the species diversity, the evenness and richness of a sample, in addition to offering a description of the sample in taxonomical terms. This method is well suited for recognizing previously studied bacteria or archaea and calculating plain phylogenetic distances. However, it can only assign meaningless etiquettes to new species.

The second strategy does not apply any such restriction, as it sequences all genetic material in the sample in a random fashion. Under the moniker shotgun-metagenomics, and by subsequently reassembling complete genomes, this method gives us the opportunity to start describing the chemical functions of the living organisms. Indeed, all reactions that a cell can perform are necessarily catalyzed by one or several proteins in its possession. Furthermore, the blueprints for the protein's fabrication are necessarily stored in the coding-regions of its genome, i.e. in the genes. Finally, using this information combined with the appropriate databases and algorithms, we can predict the metabolism of microbe populations and figure out how they contribute to the cycling of elements and energy within the system they inhabit. One should note that the responsibility and importance of microbes in planetary biochemistry is paramount in this epoch of climate change.

In conclusion, modern microbial ecology methods allow us to use a systems-based approach to ecosystem functioning, in contrast to the black box view used in classical biogeochemistry disciplines. The latter solely make use of statistical relationships between rate measurements and environmental conditions to explain functioning. However, in our opinion, without a mechanistic understanding of the underlying metabolic actors, the creation of a sufficiently accurate model for better managing and, perhaps, attempting to engineer the system, is not possible.

Read the following chapters for some example applications!

2. Introduction

2.1 The hunt through the ages

Let us not hold forth about how incredibly essential microbes are, an old mantra offered time and time again. Without them, life as we know it would not exist; thanks to them our digestive system functions, plants grow and, generally speaking, all the proper planetary biogeochemical cycles take place. Instead, we will attack our topic from another angle.

Microbes have eluded us since the dawn of time, and continue to do so today.

At first, they were unseen and unknown. Some of the processes or phenomena they caused were observable, but the effects were attributed to other forces or beings, such as mythical ones. No one guessed that the transformation of grapes into wine or the propagation of deadly pathologies might be explained by microbes. Naturally, what cannot be detected cannot be seriously considered or thought about in a scientific way. Microbes' invisibility was their first simple – but cunning – way of hiding.

Then, in the 1670s, optical technology arrived far enough to permit their visual reconnaissance and the recognition of their ontological status. The man was named Leeuwenhoek and he manufactured lenses to survey the quality of the thread in his draper's shop [3]. What he saw was a new microscopic world filled with tiny objects and small motile particles. Minds at the time were puzzled, and though the life of these *animalcules* was undeniable, their importance could not be grasped. This was the microbes' second evasive technique: their omnipresence combined with apparently aimless behaviour.

Only much later, in the 1870s, thanks to another microbe hunter, namely Koch, was a first phenomena of high public interest linked to microbes. The phenomena was that of falling sick to anthrax and the bacteria identified to be responsible for the disease was thus named *Bacillus anthracis*. The team demonstrated the causal relation clearly by growing the bacteria in a nutrient broth, injecting it into a healthy animal, and observing the development of the illness [4]. But many such crucial species of bacteria and archaea do not let themselves be tamed and domesticated in such a way.

Indeed, to illustrate this, imagine the following experiment: you are to collect a sediment sample from lake Ekoln, situated in the south of Uppsala. Mix it with sterile water, vortex it, allow it to settle, dilute the supernatant and pipette two droplets of equal size from the result. Put the first droplet under a modern

microscope and spread the second droplet on a petri dish with agar and nutrients. After leaving the dish in a warm environment for a few days, compare the results. What you will see is something resembling the schematics in figure 1.

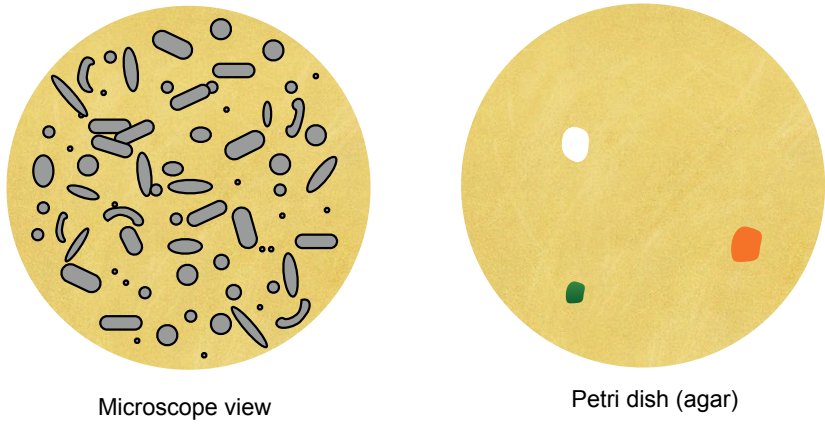

Figure 1. Schematic representation of the great plate count anomaly.

Because the two drops were of equal size, they contained virtually the same amount of microbes. Furthermore, if you assume that each single cell, once given unlimited food, will multiply in the millions and form a visible spot on the dish, this could lead you to expect that the number of colonies appearing on the agar plate should be of the same order of magnitude as the number of microbes you count under the microscope. It turns out this is not the case, the colonies are fewer by the thousands-fold. This apparent contradiction was dubbed "the great plate count anomaly" [2] and is one of the older and most profound puzzles in microbiology. The microbes' refusal to be isolated and grown in controlled environments can be seen as the third elusion they flaunt.

One of the first hypotheses aiming to explain the anomaly merely stated that the majority of cells observed under the microscope must be dead microbes. That is, individuals that are not viable any longer and are incapable of carrying out life's basic requirement: reproduction. Additionally, it was suggested that most microbes could be in a "latent" or "dormant" state [5]. These propositions were successfully falsified by several experiments measuring uptake of radioactively-labeled substrate showing that most cells were actively taking up the provided compounds [2].

Contemporary microbiology explains the anomaly by pointing to a combination of factors that all play a role in preventing the successful establishment of microbes in laboratory media. Firstly, many microbes are oligotrophic and high concentrations of nutrients and other molecules can be poisonous to them [6]. Secondly, glass vials that were used for culturing cells, and detergents

used for cleaning can be toxic and inhibit the growth of certain bacteria [6]. Thirdly, many commensalisms exists where one species produces a compound or carries out a compound that is essential for an other species [7]. Fourthly, fast growing species may overwhelm those that divide only very slowly, thus leading to an imbalance of cell-to-cell communications. In addition, inhibitory compounds may be produced that result in the inactivation of the cells by other microbes in their immediate vicinity [8].

Simply put, the niches or <habitat scope> of most microbial species are relatively narrow; they have evolved and adapted to natural conditions and the petri dish is a foreign environment that doesn't suit their life strategies. Additionally, microbial species depend heavily on one another for creating viable growth conditions. Despite the fact that often many species carry out similar functions (thus creating <switching loops> and providing resiliency to the interdependency network) removing the community from its environment creates a sharp shock causing the complete crash of the dynamic system, leaving behind only a few odd members that manage to proliferate alone.

Some partial solutions to these problems exist, we can mention here approaches such as the enrichment strategy [9] or the mixed cultures [10], without going into more details. However, on the whole, scientists have been unable to culture and study the vast majority of earth's microbial diversity [11], no matter how much the artificial conditions are tinkered with. It is estimated that only 0.1% of microorganisms from the environment have been successfully brought to the laboratory [12] [13]. The gap evidenced by the first description of the great plate anomaly in 1898 (performed by Heinrich Winterberg) seems to be just as wide as it was a century ago [14]. Scientists vain culturing efforts can amusingly and metaphorically be compared to gathering all the nutritional requirements of squirrels (nuts, water, air), mixing them in a large fermentor, adding a big chunk of forest for good measure, and hoping to culture squirrels.

Of course, we cannot ignore the microbes that do grow in laboratory conditions such as the illustrious *Escherichia coli*. These have taught us much about the functioning and organization of a bacterial cell and have been wonderful models for genetic engineering thanks to their short doubling times [15]. Sadly, they have not taught us much about ecology or the role of microbes in the environment. These laboratory microbes do not represent the relevant <players> in a natural environment where they are typically found in only very small proportions [16].

We must also note that, unlike the process of falling sick which often depends on a single species of microbes, the ecosystem processes we aim to study do not depend on single groups of microbes. Even if these groups were readily obtainable, they could not be directly used to answer the ecological questions at hand. This is in part what has lead scientists to develop new techniques that capture snapshots of whole microbe communities in lieu of focusing on individuals and clonal populations.

Enter modern molecular methods! At last, we can move forward to the 2010s in our story and use the latest artillery in our pursuit to pierce the stealthiness of our opponents. Nowadays, the widespread and popular approach for solving the microbial puzzle is to grasp at them via genetics. By attempting to measure the content of their genes, we can start to better hypothesize on their functions and organization. In this way, the culturability issues are bypassed and all that is required is a straight-forward total-DNA extraction protocol on collected samples followed by the use of a sequencing machine. Happily, since the beginning of the millennia, the cost of applying such molecular methods has dropped drastically, as seen in figure 2.

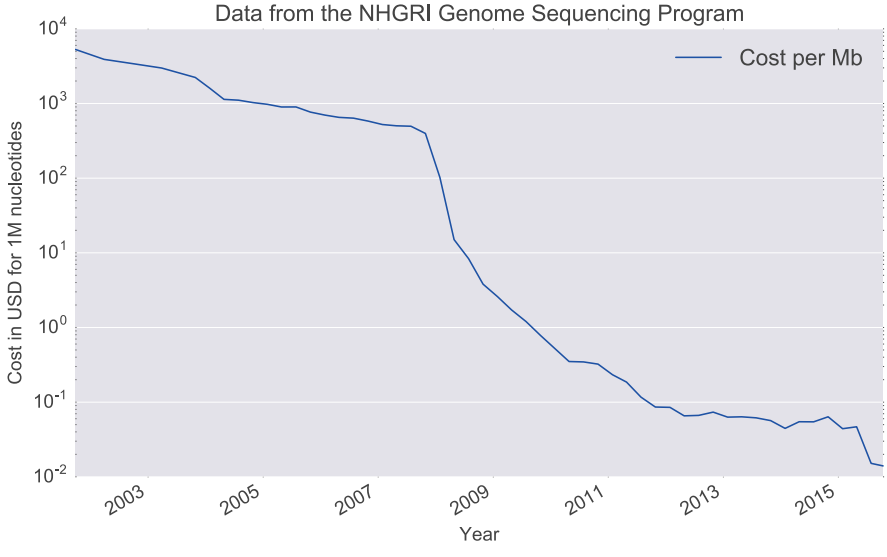

Figure 2. Cost for sequencing a million nucleotides over time.

By collecting microbial assemblages directly from the environment and subjecting them to high-throughput sequencing, a mixed and fragmented semi-random subset of the genetic content of all microbes inhabiting the sample is retrieved. The quantity of data produced is very large and its structure is complex; extracting meaningful information from it requires great skill. Naturally, the drawback is that the analytic procedures implemented create a whole new set of problems. On the bright side, the data produced can help the scientist to solve some mysteries. For example, we can now build better phylogenies between microbes [17], resolve community composition in terms of taxonomic units [18], inspect genes and metabolic pathways [19] [20] as well as compute statistics on strain variations [21]

We will delve into more details about dealing with microbial sequencing data in the methods chapter and individual papers further on.

To conclude, we see that the hunt through the ages has moved to an entirely new battlefield. From the optical challenges early on, to the physical and biological challenges that followed, we have arrived at the informational challenges of the hunt for capturing and understanding of microbes. We are now chasing them in the digital realm and our lenses or incubators have become computers and databases. Microbiologists must double as bioinformaticians, though university curricula have been extremely slow to adapt. Sometimes I wonder how Pasteur would feel about programming python analysis scripts and am not sure about his reaction.

Nevertheless, microbes continue to divert our efforts in what can be considered the fourth set of obstacles presented: They defy the construction of our phylogenetic trees by exchanging genetic material horizontally (e.g. a trait inherited not by paternity, but by physical proximity between brothers). They ridicule our genome reconstructions by rapidly changing the <flexible> part of their genetic repertoire [21] [22]. They hide from our compositional studies by possessing widely varying ribosomal genes or by clustering too closely with a similar species [23]. They dissimulate their diversity by camouflaging within the errors and uncertainties of the sequencing machine [24].

They laugh as we are still trying to get ahold of them with our giant and slippery hands while their secrets remain safe. But despite their furtiveness, they will be cornered by human inventiveness one day...

2.2 Definitions

We have proceeded so far without defining the word <microbe> and are in dire need of precision. The simplest definition is often merely based on size: A microbe is a reproductive organism smaller than 100 µm, i.e. the tenth of a millimeter.

To picture it, think of anything of equal or smaller size as compared to the black speck situated on this line of next, directly after the colon: .

However, such a simple classification criteria would include viruses on the lower spectrum, as well as complex multicellular organsims on the higher end, neither of which are of interest in our studies. In this thesis, we shall use a slightly more restrictive definition:

• **microbe** /mʌɪkrəʊb/ *noun*
A single-celled microorganism.

We introduced the term bioinformatics earlier, a new field of science that has emerged in the last decades, not specific only to microbiology. Bioinformatics is often taken to mean “biological computer science” but that is incorrect. Here is what I hold it to be:

• **bioinformatics** /ˌbʌɪəʊˌɪnfəˈmatɪks/ *noun*
is the science and technology of biological information. It is a theoretical science that develops mathematical, statistical, and algorithmic methods and makes use of experimental data (in digital form), in an effort to discover functions and modalities of information processing in the cell.

The thing is, the management, storage and processing of biological data across large computer networks is an important, albeit technical, aspect of bioinformatics and is often presented as bioinformatics itself, which it is not. This means for instance that when you are writing a program to convert the input file from one format to another because the analysis program you want to use doesn't accept the XYZ format, you are not doing science, you are more akin to a data janitor at that moment.

I like to say bioinformatics is concerned with modeling and studying biological processes not as chemical processes nor as physical ones, but as informational processes. For instance, if we consider the differences between DNA and mRNA, a chemist could describe DNA in a cell as chromatin and would view mRNA as smaller more fragile molecules. A bioinformatician, on the other hand, would say DNA is a constant piece of information in the cell while mRNA is a rapidly changing one that is derived from the first, perhaps, in analogy to a hard disk and the RAM.

Next, the method introduced above, where one collects an environmental sample of microbes – such as a gram of soil or a glass of lake water – and processes it in a high-throughput sequencer to retrieve a snapshot of the genetic content of the inhabitants, has acquired the moniker <metagenomics>. The *omics* suffix conveys the notion of a systematic and comprehensive study, a sort of totality. Contrasted with the more usual term <genetics>, one could say that genomics is the simultaneous study of all genes, all discrete features in a genome. The *meta* prefix suggests a supplementary level of encompassivness and signifies that we are studying more than one genome. In fact, a large aggregate of genomes.

• **metagenomics** /ˈmɛtədʒɪˈnəʊmɪks/ *noun (uncountable)*
is the study of genetic material recovered directly and in bulk from environmental samples. It is notably divided into two distinct procedures: targeted-metagenomics and shotgun-metagenomics.

Lastly, let's describe a useful ecological term that will be used many times across this thesis:

• **trait** /ˈtreɪt/ *noun*
A trait is defined here as a phenotypic characteristic of an organism that is linked with its fitness or performance [25]. This includes anything directly measurable such as the color of an organism, its ability to move, its toxicity or even features such as genome size.

2.3 Microbial biogeography

After describing the history of microbial hunters and defining terms, let us turn to some concrete scientific questions and consider what type of contributions the new era of microbiology can provide.

The new molecular tools available have radically changed our way of thinking about microbial evolution and upset the hierarchical taxonomic structure that Linnaeus had proposed. Spurred by the studies of Carl Woese [26] and his colleagues [17], the microbial species dilemma has been brought forward and, to date, no concept allowing a satisfactory definition of a microbial species has emerged [27]. The rules that apply to sexually reproducing mammals do not translate well to dividing cells that exchange genetic material frequently and horizontally. I would like to leave the impossible quest for a fundamental taxonomic unit of biological classification (i.e. a species) behind and instead imagine a simple continuous population landscape forming peaks or clusters of evolutionarily linked microbes across its surface, as illustrated schematically in figure 3. The x-axis of this figure could of course be expanded to N-dimensions and the qualifiers could be changed from arbitrary species names to functions or traits.

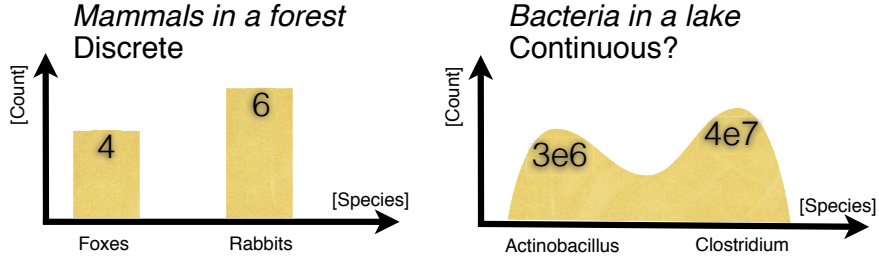

Figure 3. Discrete or continuous units of biological classification.

Yet, we may rely on an operational definition of a taxonomic unit that approximates sets of closely related microorganisms in order to describe some properties of microbial communities such as variations in diversity across space or time. This can include the number of peaks in the nucleotide polymorphism landscape which can be used to compute diversity indicators. The last two decades of microbial biogeographic studies based on operational definitions of microbial taxonomic units have shown that, like plant and animal distributions, microbial distributions are the result of both deterministic and stochastic processes [28].

A long-held ecological concept is the following: microorganisms are dispersed globally and are able to proliferate in any habitat with suitable environmental conditions. This concept was introduced by Martinus Beijerinck and concisely summarized by Baas Becking in the widely referenced quote

“everything is everywhere, but the environment selects” [29]. In other words, the only limiting factor for establishment of a species X in location Y is the global properties of the habitat at hand – dispersal forces are considered infinite. The rationale was that the small size and high abundance of microbes (as well as other aspects of their biology) increase the rate and scale of geographic dispersal to levels where any limitations are nonexistent, resulting in <cosmopolitan> distributions. However, recent application of molecular methods has challenged the conceptual dogma, providing evidence of microbial endemism [30], and also of a spatial patterning of microbial biodiversity [31] [32] qualitatively similar to that of plants and animals.

As with macroorganismal biogeography, microbial biogeography initially adopted a taxonomic approach, focusing on genetic signatures to identify groups of microorganisms. These studies revealed classic patterns such as the species-area relationship and isolation by distance [28]. First, large efforts to survey the global microbial diversity revealed the distribution of DNA sequence clusters in earth’s major biomes such as the ocean [20], leading to the definition of the habitat ranges of phylogenetic branches of the tree of life. Furthermore, these studies suggested the existence of latitudinal patterns of biodiversity with microbial diversity decreasing towards the poles [33], as observed for macroorganisms [34]. Large strides forward have been made in understanding the coupling between local and regional scales using the metacommunity concept as a guiding framework [28]. However, microbial biogeography along the environmental spectrum ranging from soils to inland waters and the ocean was not well understood until recently. A conceptual framework integrating the biogeography of microorganisms over different kinds of ecosystems is the River Continuum Concept [35]: it assigns high importance to the biome interfaces, to the physical drivers such as water flow, and to wetted perimeter in determining diversity patterns at the landscape scale.

Another question that microbial ecologists have been occupied with for at least half a century is the famous <plankton paradox> [36]. The paradox is based on the following assumptions and observations. It was first thought that the diversity of the microscopic plankton inhabiting the oceans and lakes should be quite low, due to the perceived uniformity of conditions across large areas of water bodies and the effect of the ecological principle known as competitive exclusion. This principle states that when two species compete for the same resource within the same habitat, over time, one will necessarily be driven to extinction. The paradox is that while the range of resources to compete for in an ocean is relatively limited, the observed diversity of microbes is immense.

Even though water seems uniform to the naked eye, diverse niches and micro-environments become obvious when looking through a microscope. At present, the plankton paradox is explained by demonstrating that the microbial habitats never reach equilibrium [37]. The solutions put forward by modern microbiology are to introduce the following aspects to the model:

- Vertical gradients of resources. The distribution of energy sources is not the same at the top and at the bottom of the water column.
- Differential predation. Larger animals that eat microbes will influence which species thrive and which ones are driven to extinction.
- Micro-environmental structure and turbulence. For example, small elements such a pellet of food moving under Brownian motion will create local <hotspots> of activity.

The second long-held ecological concept, still widely debated, concerns the relationship between microbial biodiversity and ecosystem stability. The stability of a biome can be defined as its resistance and resilience to disturbances and its stability can be investigated in terms of functional or compositional parameters. As microbial communities are highly diverse and include many functional redundancies across its members, microbial systems are expected to be enduring in terms of functional parameters when subjected to a disturbance. The existence of rich and abundant seed banks [38] and low-dispersal limitation [39] purportedly provide ecosystem stability. Nevertheless, stability may depend in large part on the particular function of interest [40]. For functions that are carried out by many taxa [40], a high degree of functional redundancy exists [41] and changes in community composition or loss of taxa may not correspond with changes in functional rates [42]. Alternatively, for functions performed by only a few taxa, the sensitivity and resilience of the function may closely follow changes in the abundance of those taxa. Notably, estimates of resistance and resilience for the same microbial community may have different values depending on whether compositional or functional responses are measured and on which functions are used to assess stability.

However, interpreting taxonomic patterns in terms of how they are related to the function of a population or community is especially difficult in microorganisms, where a broad range of functional variation may occur among similar organisms (e.g., organisms with the same 16S rRNA sequence). Thus, there is growing interest in the biogeography of functional traits in microbial ecology in a similar fashion to macroorganisms [25]. In plants and animals, focusing on traits has greatly helped in studies concerned with describing mechanisms underlying community assembly [43], studies pertaining to the relationship between biodiversity and ecosystem functioning [44] as well as descriptions of an ecosystem's response to environmental change [45]. While several authors have proposed using trait approaches in microbial ecology [32] [46] [47] [48] [49], the applicability of these approaches is hampered by methodological constraints. Initial attempts include the study of patterns in genome size [50] as genome size is thought to be indicative of organismal life strategies [51]), the degradation potential of communities using biolog plates [52] and shotgun-metagenomics studies [20].

Shotgun-metagenomics can provide detailed insights into the DNA blueprint of the community, in particular concerning the genes encoding the ecosystem-function determining traits. However, technical issues such as the quantifi-

cation of traits from shotgun-metagenomic data and the partitioning of traits among taxonomic groups and populations need to be resolved. For the quantification of traits, various standardization procedures have been proposed to determine genome equivalent estimates, an approximation for the proportion of genomes harboring a specific gene [53]. For the partitioning of traits, assembly and contig binning approaches have been developed [54] providing some first high quality genomes of previously uncultured deep branching clades [55] [56]. Additionally, a classification system for traits is an important prerequisite for functional traits based biogeography: first attempts have been published [49]. Besides these methodological challenges, conceptual frameworks must be adapted to the new data, so as to provide a solid theoretical basis.

2.4 Metagenomics

Metagenomics applies a series of genomic technologies and bioinformatic tools to access the genetic content of entire microbial communities. Over the past two decades, the field of metagenomics has been responsible for substantial advances in microbial ecology, evolution, and diversity. Indeed, many research laboratories are now actively using metagenomics methods [57]. There are two main approaches, namely targeted-metagenomics and shotgun-metagenomics.

Until now, we have used the term "targeted-metagenomics" without defining what the actual *target* is going to be. The idea is to target a specific gene or gene region that is found in all microbes with only minor differences between species. This is the only way to design an experiment that can capture the total diversity contained in a sample. For instance, if a gene that participates in photosynthesis is selected as the target, only diversity in phototrophic microbes will be measured, and all the other microbes that derive their energy in different ways will be ignored.

To satisfy their energetic needs, all living cells catalyze chemical reactions. To catalyze reactions, cells make proteins, and, to make proteins, cells need some apparatus that assembles the correct chain of amino acids by using information contained in their DNA. This is the role of the ribosome. Absolutely all living cells seen on our planet to date had ribosomes in them. From this fact we can conclude that all microbes necessarily have the genes to make ribosomes stored somewhere in their genome: We can use this property to our advantage and target one of the particular genes that participates in this process to probe a microbial population. Concerning bacteria, the most used and adopted standard in microbiology today is to target one or several regions of the <16S> subunit gene of the ribosome. This gene acquired its name from early experiments performed to separate the various components of the ribosome, cell lysates were ultracentrifuged and that particular RNA subunit was always found at the 16 Svedberg sedimentation mark (non-SI unit).

The 16S rRNA gene is particularly well suited for diversity estimates as it contains highly conserved parts (low variation between species) that can be used as annealing regions for the forward and reverse primers, interrupted by highly variable parts [58]. These highly polymorphic parts have not been under strong selection pressures and were able to <drift> randomly across evolutionary time scales, presenting sequences unique to each species or even unique to each strain of bacteria. In our studies, we targeted the V3 and V4 variable regions. Most often, the variable loci describe regions of the ribosome that do not participate in three dimensional conformation. For example, the nucleotides involved in RNA hair-pin structures are less likely to display polymorphisms as a mutation at that locus has a very high chance of being deleterious and resulting in a dysfunctional ribosome [58].

To summarize previous accomplishments, targeted-metagenomics approaches, when combined with HTS, have provided a detailed picture of microbial diversity around the globe, from environments of the arctic to antarctic [59], the deep ocean [60] and freshwaters [61]. This deep sequencing has allowed the detection of uncultured microbes including even the very rare parts of the microbial life inhabiting our planet.

The other metagenomics approach, shotgun-metagenomics, enables the researcher to comprehensively sample a random subset of all genes in all organisms collected from the environment. By using this method, he can begin looking at all the parts of the genomes and in particular those parts that relate to function such as metabolism. Using such data, different types of analyses have been applied including genome-centric and gene-centric approaches. A gene-centric analysis treats the community as an aggregate, largely ignoring the contribution of individual species and genes and gene fragments from a given metagenomic dataset are mapped to gene families, providing an estimate of relative representation. The power of the approach lies in comparing relative gene family or subsystem abundances between metagenomes to highlight functional differences [62].

By using a genome-centric approach, the scientist attempts to associate his data with the organisms from which they were derived. The process of association between sequence data and contributing higher-level taxonomic groups is called binning or classification. The most reliable form of binning is assembly. That is, in a perfect assembly, all reads in a contig are derived from a clonal population, with the optimal binning being a closed circular contig. However, this is often not the case, and some level of co-assembly is usually encountered in metagenomic derived genomes, particularly as there are almost no clonal microbial populations present in nature. While different binning methods have been proposed to overcome some of the limitations of assembly, this genome-centric approach rarely provides the precision required for discriminating between closely related strains.

Still, such a coarse level of sequence assignment can be useful for interpreting microbial communities. These less stringent assemblies, which certainly

produce chimeric and misassembled contigs, can provide useful insights into previously uncultured microbes [55] [22]. One of the first success stories of the shotgun-metagenomics sampling technique was a paper by Tyson and colleagues [19] which sampled self-sustaining biofilms growing on the surface of the hot water draining out of an abandoned mine. Thanks to the low number of species living in the biofilm, the authors managed to fully reconstruct the genome of two of the dominant bacterial species. Tyson et al. was the first publication to ever achieve reconstruction of multiple genomes directly from a natural sample. Specifically, this investigation was able contribute to answering difficult questions such as: "How do these communities resist the toxic metals of the mine drainage and maintain a neutral cytoplasm?"

Another study cataloged enzymes and taxonomic groups in the open ocean [20] while others compared the protein space between marine and freshwater ecosystems [50]. These and other shotgun-metagenomic studies revealed the metabolic potential residing in aquatic environments and lead to inferences concerning the traits driving ecosystems functions. It can be argued that directly assessing the functional diversity encoded in the genomic content of a community is much more accurate than indirectly assigning metabolic potential to a community through phylogenetic relatedness as obtained by targeted-metagenomics. Considering that certain traits can be horizontally transferred between organisms (whether distantly or closely related), shotgun-metagenomics provides the more reliable predictions.

Yet many caveats and pitfalls remain when performing shotgun-metagenomics. To master shotgun-metagenomics, a wide series of different technologies must be applied, including techniques from advanced molecular biology and bioinformatics. We assume that the extraction of the genetic material will take place in equal amounts and that the sequencing will be performed in an entirely random fashion, but these steps cannot possibly be carried out without at least introducing slight biases. The choice of an assembly algorithm and binning strategy (i.e. differential abundance, nucleotide composition or taxonomic relatedness) may all result in different assembled genomes effecting downstream functional and taxonomic predictions. Lastly, the choice of the homology search algorithm to run and databases will affect the final annotation outcome.

Another interesting aspect of metagenomics concerns scientific deontology. As these new molecular tools are able to measure so many variables simultaneously and new modes of carrying out microbial ecology are developed, many scientists have started to favor explorative over hypothesis-driven surveys. This contrast can be seen by opposing the two project descriptions "I would like to know if process X, followed by the manifestation of Y causing Z does or not occurs, in this order, in this lake" and "Let's sequence two hundred batches of microbial filters along a timeseries in this lake and see if we find something interesting after computing all possible correlations and maybe skipping the application of Bonferroni correction".

This effect is also caused, in part, by the many methodological challenges that need to be overcome, absorbing all the attention of the students or scientists and leaving the experimental design in second place. Though most studies are not guided by ecological theory, there are examples of hypothesis-first publications, such as a very recent study of the gut microbiome [63]. There, the authors test the idea that multifunctional redundancy is an intrinsic property in the gut reflecting the health of the human ecosystem. Another examples include: (i) a survey of uncultured Archaea from deep sea sediment providing answers to the debate surrounding the origin of eukaryotes. A clear case was made for the eukaryotes representing a branch within the archaeal domain [64]. (ii) A time series of shotgun-metagenomes from a freshwater lake demonstrated that different models of bacterial speciation can be applied to different populations coexisting in the same environment [22].

To summarize, there have been high expectations concerning the topics and problems that metagenomics would provide answers to, such as resolving the microbial species challenge, detecting elusive functions in new ecosystems, capturing the unknown branches of the microbial tree of life and thus providing insights in this so called <microbial dark matter>. To move forward, the metagenomic analysis toolset must be expanded and current methods need to be validated.

2.5 Reverse ecology

The term "reverse ecology" was first coined by Matthew Rockman at a conference on Ecological Genomics [65] . Rockman describes the study of organisms and/or processes in a bottom-up approach, which takes off with the genetic information of the environment, the entities, of whatever kind, under study. In "Advances in Experimental Medicine and Biology" [66], the central concept of reverse ecology is illustrated by contrasting the studies performed by Darwin on the Galapagos islands with the modern DNA-sequencing experiments of today. In the 1800s, Darwin observed the different finch species in the Galapagos islands: he rigorously described the different beaks of the different species. Darwin explained the evolution of the differing morphologies as linked to the different food sources available and the competitive-exclusion principle. Darwin's observations were phenotypical, and were only linked to underlying genotypic causes over a century later. Namely, to the set of genes expressing signaling factors that play a role in the development of finch beak morphology, as presented in [67]. *Forward ecology* shows how a specific environment causes a variation in phenotypic traits, and that modern advances can delve into the genetic basis as one of the primary causal mechanism of these differences.

In reverse ecology, the scientist starts off with the genetic information obtained from an organism or an environmental sample and attempts to predict

the identities, functions, phenotypes of the community or individual cells inhabiting the environment under study. Using the nucleotide information without any *a priori* assumptions, hypotheses and conjectures are formulated, with the aim of determining the traits or processes that play a role in the organisms under scrutiny, as well, of course, as their relationships to their environment. A well-known example of the approach and method is the detection of proteorhodopsin, a light driven proton pump, revealed by shotgun-metagenomics as acting, originally, in an uncultured bacterial taxon [68]. Later, homologous were found in most bacteria inhabiting the surface ocean [20]. From this, we can likely assume that these microbes can use sunlight for energy inputs. Provided these genes make active catalysts, this discovery can profoundly modified the comprehension of carbon and energy flow in aquatic environments [69].

By using homology inferences, Rusch et al. evidenced that most bacteria can use sunlight for energy inputs. Provided these genes make active catalysts, this discovery can profoundly modified the comprehension of carbon and energy flow in aquatic environments [69].

Pursuing the path from genotype to phenotype is the defining approach of reverse ecology studies and experiments, similar to reverse engineering, which determines, when possible, the role of each of the constituent, individual parts, of a machine, object or analytic process, contributing to the functioning of the whole. In reverse ecology, reconstructing the functioning of complex communities present in natural environments is a real possibility. Taking an example presented later in this thesis, the genomic content of boreal lakes indicates high potential of inorganic carbon fixation demonstrating that the million of boreal lakes not only process carbon from the surrounding landscape but also that many organisms can incorporate carbon into organic matter. A complex network of metabolic reactions with the potential of complete carbon turnover emerges where genome reconstructions revealed a wide range of trophic strategies present in the bacteria inhabiting boreal lakes. Annotating the genomic content can also be extended to study other elemental cycles such as those for S, Fe, N and P dominated by biological activity.

Most chemical reactions that happen passively are slow, except for instance at the oxic-anoxic interface of the biosphere. The reactions that take place and sustain life are those that are catalyzed by nanobiological machineries composed largely of multimeric protein complexes associated with a small number of prosthetic groups [70]. The blueprints for these protein complexes are imprinted in all the genomes of the diverse organisms inhabiting our planet. As such, reverse ecology could even be extended to the entire globe. To summarize reverse ecology as it is applied to microbial ecology, it attempts to place the genetic elements of a community in a systems biology framework, in order to create a model which may serve to predict the behavior of the ecosystem.

Another goal of reverse ecology is to form a network useful for predicting the interactions between organisms [71] as well as understanding why the genomic

content is assorted in the way it is on the local and regional scales. Hopefully, reverse ecology contributes to unravelling the principles that control the adaptation of cellular life to its environment. In other words, a fundamental goal of reverse ecology is to transform genetic information into ecological knowledge. Naturally, the fact that the entities present in an environment have adapted to the environment over evolutionary time scales must be kept in mind.

A large part of the genomic content of an organism (particularly for Bacteria and Archaea) is devoted to coding for proteins which in turn catalyze specific chemical reactions and define the potential metabolism accessible to an organism. Freilich and colleagues [71] analyzed the link between ecological strategies and growth rate of 113 bacterial species, with full genome sequences. Their results, obtained by metabolic-network analysis, strongly support the dual specialist vs. generalist groups concept. Moreover, they unveiled an underlying general principle: the growth rate of different microbes was explained by their metabolic potential. They found that larger genomes are linked to faster growth rates, and that faster growth rates are linked to species with greater ecological diversity (larger metabolic networks). For instance, the fastest growing group identified was the dual aerobic and anaerobic bacteria. By applying their <seed set> analytical framework to the genomic data they rebuilt the probable environment for each microbe. Freilich et al. put forward a reasonable explanatory factor, i.e. that lower competitive pressure is exerted on very specialized organisms, which in turn implies, for these organisms, lower growth rates.

Reverse ecology in metagenomics has, so far, been usually implemented to describe organisms and their traits, as well as the metabolic potential of ecosystems, which is a pity. Few inferences or conjectures concerning general ecological principles have arisen, despite the fact that the reverse ecological approach, combined with metagenomic data, offers vital information crucial for addressing long-standing questions in ecology. How are ecosystem processes determined? How to describe, delimit, the stability of a microbial system? Are species diversity and composition the indispensable descriptors which must be taken into account? Both have been judged as significantly affecting many ecosystem processes. Yet, traits and organismal interactions may be community features more essential for determining the productivity, elemental cycling and system stability [44] [45] [49].

Despite consequent computational advances, the analysis of community genomics data within a meaningful ecological framework remains an elusive goal [72] [62]. Metagenomics produces data suitable understanding and charting ecology. Though it allows for performing “high-throughput ecology”, the scientific community has not bridged the gap between theoretical and analytical tools to unveil genomic community patterns and the processes that underlie them, ultimately determining ecosystem processes.

3. Aims

The aim of this thesis is to study several aspects of microbial ecology with the use of the modern molecular tools that are described by the over-arching term "metagenomic sequencing". First, we will develop novel bioinformatic tools (or novel ways of combining existing tools) and, secondly, we will apply them to various datasets in an effort to gain insight into the spatial organization and distribution of microbes, their diversity, their interactions and their metabolic functions. Additionally, this knowledge is to be contextualized and placed in an environmental framework where microbial activity can be associated with the flux of energy and matter, ideally producing a mechanistic understanding of the underlying processes and enabling us to improve our predictions of ecosystem responses to changes in the bioshpere. The variations and techniques of the wet laboratory methods are set aside; our focus in this thesis is on the bioinformatics methods involved.

We chose inland water systems as models for our studies. The datasets examined in this thesis were all collected and generated from lentic and lotic water bodies including: alkaline lakes of Austria, the Danube river and humic lakes of the Swedish boreal landscape.

More specifically, these topics were raised and the following projects accomplished:

- In paper I, we planned to design and validate a targeted-metagenomics protocol that would be reliable and able to process sufficiently large number of samples in parallel to adequately address the scientific questions we had in mind (at least hundreds of sample simultaneously).
- In paper II, we wanted to make use of online databases and the publicly available information to add context to our own datasets by performing sequence similarity searches and carrying out text mining on the results.
- In paper III, we used the two methods developed and the expertise built and applied them to a large riverine system. Our aim was to study biogeography patterns of bacterioplankton and to lay the foundation for a unified concept of microbial community assembly linking terrestrial and lotic systems.
- In paper IV, we set out to study the microbial metabolic networks in stratified boreal lakes. By reconstituting the metabolism along several depth profiles, we tried to understand the distribution and partitioning of pathways and reactions ultimately determining green-house gas emissions from theses systems.

- In paper V, we attempted to reassemble the composite genomes of the abundant microbial inhabitants of alkaline soda lakes, as well as to describe the organisation of microbes and their metabolic strategies under alkalophilic and psychrophilic conditions. In particular, we focused on the adaptations related to inorganic carbon uptake and the enzymes involved in polymer-degradations (i.e. hydrolases) under different pH conditions.

Overall, this thesis contributes to the field of microbial ecology in the areas of: HTS data processing for microbial species and diversity analysis, the automation of relevant information extraction from databases, the understanding of biogeography patterns of bacterioplankton, the distribution and partitioning of metabolic traits responsible for the production of green-house gases in the boreal landscape and, finally, the metabolic adaptations of microbes in high pH and low temperature conditions.

4. Methods

4.1 Overview

All the studies in this thesis have in common the fact that they make use of metagenomics techniques (for a definition see section 2.2), with variations in the laboratory protocols and data analysis. There are two main types of metagenomics: "targeted-metagenomics" and "shotgun-metagenomics".

Briefly, to perform metagenomics, you have to first carry out the following operations in the physical world:

- You collect a water sample (or a sediment sample as in paper I) from your ecosystem of study. The quantity of water depends foremost on the number of microbes per millilitre. An average value of one liter can be used.
- You filter the liquid through a sieve with very small pores and throw away what is retained. This step will determine what kind of organisms will be included in the results and stands as your practical definition for what is a microbe. Typically pores of 100 μm. Some microbes, such as the filamentous forms of bacteria that grow in multicellular colonies may be filtered out. In addition, you use a second filter with even smaller pore sizes and throw away what passes through. This will discard the viruses and define the cutoff for what is not a microbe. Typically pores of 0.2 μm.
- Then, you extract all the DNA that was contained in the cells collected. This can be done in several ways, such as bead beating, ultrasound sonication, electrolysis or chemical lysis. At this step you also fragment the long strands of DNA into shorter pieces that measure, on average, approximately 10'000 base pairs.
- After a few more laboratory manipulations such as amplifying the DNA with a PCR and quantifying it, it is ready to be delivered to the sequencing facility and inputted, with the correct reagents, into the high-throughput sequencing machine. The type of machines used in this thesis all work by the same principle of parallel and step-wise complementary strand synthesis using nucleotides with fluorescent dyes attached. Such that, at every round of synthesis, a particular light frequency is emitted and recorded by a camera. Typically A is red, G is yellow, C is blue and T is green. The images taken are analyzed automatically to recreate the original DNA sequences with high certainty.
- About two or three months later, when the sequencing facility finally gets to your position in the queue, you receive a bunch of files placed somewhere on a server. Usually, two files per original biological sample.

A raw view of the type of data produced can be seen in figure 4. Such a file typically contains millions of sequences, however, here in the small window of the text editor, one is only able to see the first four.

Figure 4. Viewing a FASTA formated sequence file in a plain text editor.

It's interesting to consider that, with the current sequencing technology, the number of <reads> (i.e. DNA sequences) you get is roughly on the same order of magnitude as the number of microbes in the original sample.

Of course viewing the result like this doesn't serve much purpose and you are unable to say anything about the microbes that were living in the environment you sampled. You need to process it and you can't to it by hand. You are going to need a series of programs and tools to extract biological meaning from this mega- to terabytes of sequence data. So you step away from the physical world and must now manipulate the data in the digital world, typically by following these steps:

- *Demultiplexing*: often, sequences from multiple samples are mixed into one vial to increase the through-put of the sequencing. Before they are combined, a custom barcode, i.e. a short oligonucleotide sequence of 5 to 8 base pairs is ligated to one or both ends of every sequence. This enables you to later pick out which sequences were part of which samples.
- *Quality checks*: to assure that everything went well during the sequencing, you graph different statistics such as the distribution of sequence lengths, the amount A and Ts against the amount of G and Cs.
- *Cleaning*: as the sequencing machine provides a quality score – for each nucleotide it emits – which represents the certainty of having measured the correct base, you are able to filter and remove sequences or parts of sequences which are deemed too low quality. Typically using a average-based sliding window strategy.
- *Assemble*: this part will be very different depending on whether you are performing targeted-metagenomics or whether you are performing the alternate technique of shotgun-metagenomics. Though both methods will include a type of assembly or <joining> step where the reads are regrouped to form longer sequences.
- *Similarity search*: once again, the specifics of this part will depend heavily on what type of data was produced, but almost every study includes the use of the context and information built by the scientific data over the last 30 years in the form of online databases in a effort to make sense of newly produced datasets. The simplest approach is often to compute a similarity search between one's own data and the public sequences. Because of the varying completness and quality of the databases, these results must always be taken with a grain of salt.
- *Graphing*: there is no one unique procedure here, and what is to be done depends entirely on the scientific questions at hand, but all projects will need to produce graphical visualizations that summarize and enable a holistic view of the underlying data.
- *Statistics*: in addition to the visualizations, you will most likely want to establish some statistical tests that will enable you to validate or refute the original hypotheses that were formulated.
- *Automate*: What happens if you realize, after viewing the final statistics, that one of the barcode sequences used in the demultiplexing step had one nucleotide mistyped? Or imagine that, suddenly, in light of new developments, the cleaning procedure is to be made more stringent? Are you to manually rerun the whole analysis? This would make such studies too time consuming and impossible to realize. All the steps above must be automated, such that, after correcting the typo in the barcode, a simple command issued is sufficient to re-generate all the visualization and p-values. This is one of the most challenging parts of the metagenomics method for biologists not having received the proper IT training.

Further details on the algorithms and tools used to analyze the data presented in this thesis are presented in the individual manuscripts and will not be repeated here.

4.2 Samples and library preparation procedure

Paper II used samples collected by third parties and borrowed for the purpose of testing the developed method. Those samples will not be discussed here. Otherwise, the fabrication of samples was fairly similar in all papers. Water was captured straight from the natural environment and filtered through 0.2 µm filters on which microbes were collected. To allow aggregated microbes to be included and as no macro-organisms where observed in the water, no pre-filtering was used. There are two exceptions to the above statements:

- In paper I, a sediment sample was additionally included. This sample was taken from a sediment core incubated at 21 °C [73].
- In paper III, two separate fractions were taken successively. First using a 3.0 µm filter, followed by a 0.2 µm filter.

After a fixed amount of liquid had passed through the filters, they were removed and frozen at -80 °C or put into liquid nitrogen. The filters remained at low temperature for storage until further processing. Once back at the laboratory, the PowerSoil DNA Isolation Kit sold by the MO BIO company was used. The protocol described by the manufacturer was followed to produce 50 µl containing pure DNA before moving on to the amplification step or library preparation.

In the case of shotgun-metagenomics, standard protocols for library generation were followed. In particular a Thruplex FD Prep kit from Rubicon Genomics was used according to the manufactures protocol (manuel QAM-094-002).

In the case of targeted-metagenomics, a two-step PCR was performed according to specifications seen in tables 1 and 2.

Step	**Temp [Celsius]**	**Time [seconds]**
Initial Denaturation	98	30
20 cycles	98 62 72	10 30 30/kb
Final Extension	72	120
Hold	6	∞

Table 1. Time and temperature cycles for the frist PCR step.

Between the two PCRs, a purification step was performed to eliminate loose primers and eventual primer dimers or unspecific products of low molecular

Step	Temp [Celsius]	Time [seconds]
Initial Denaturation	98	30
15 cycles	98 66 72	10 30 30/kb
Final Extension	72	120
Hold	6	∞

Table 2. Time and temperature cycles for the second PCR step.

weight. This cleaning was executed with magnetic beads (Agencourt AMPure) following the manufacturer's instructions. After the PCRs are done, quantification of DNA concentration was realized using a PicoGreen assay (Quant-iT PicoGreen, Invitrogen). To measure the average fragment length a BioAnalyzer (Agilent) was used.

There are two exceptions to the above statements:

- In papers I, III, and V, samples were grouped into pools and double 5 bp barcodes were manually added in an extra amplification step. For each pool, the library preparation was performed separately following the TruSeq Sample Preparation Kit V2 protocol with the exception of the initial fragmentation and size selection procedures.
- In paper IV, to avoid the three-step library preparation scheme, Illumina adapters with barcodes were integrated into the primer pairs. Thus only a two-step PCR was required and resulted in a complete construct ready for sequencing.

In the both cases, the ends of the final constructs were as so (handles-barcode-adapter):

Forward: `3'-AATGATACGGCGACCACCGAGATCTACAC-[i5 index]-ACACTCTTTCCCTACACGACG-5'`

Reverse: `3'-CAAGCAGAAGACGGCATACGAGAT-[i7 index]-GTGAC-TGGAGTTCAGACGTGTGCTCTTCCGATCT-5'`

Each library amounted to a total volume of approximately 25 to 50 µl and a concentration of 5 ng of DNA per µl. Finally, samples were ready to be handled by the sequencing machines.

Sequencing was carried out using a combination of two versions of the Illumina platform:

- The Illumina MiSeq series in paper I, III, IV and V.
- The Illumina HiSeq series in papers IV and V.

The routine "TruSeq" protocol was applied by the SNP&SEQ sequencing facility in Uppsala, except in paper I, where we additionally used data coming from a previous generation "454 Life Science" pyrosequencing machine (Roche).

4.3 Data

The amounts of data were quite large. A brief summary of the number of unique sequences generated is shown below, for each study:

- Paper I: 10'338'568 reads of 2x150 bp.
- Paper II: n/a.
- Paper III: 1'572'361 reads of 2x150 bp.
- Paper IV: 6'462'674 reads of 2x300 bp and 716'818'649 reads of 2x100 bp.
- Paper V: 928'504 reads of 2x250 bp and 664'914'151 reads of 2x100 bp.

Several computer clusters were used to store, process, analyze and compute statistics on the data for each study.

- Paper I: The <kalkyl> node cluster from the UPPMAX facility at Uppsala university.
- Paper II: The <quince-srv2> high-performance server from the university of Glasgow in the United Kingdom.
- Paper III: The <milou> node cluster from the UPPMAX facility at Uppsala university.
- Paper IV: The <taito> node cluster and the <sisu> cray of the CSC computing resources at the university of Jyväskylä in Finland.
- Paper V: The <ww-hmem02> high-performance server on the CLIMB network via the university of Warwick in the United Kingdom.

4.4 Other parameters

In many of the studies, numerous physical and chemical parameters were measured. A non-exhaustive list of the methods used is given below:

- Water temperature and oxygen concentrations were measured in situ using combined temperature and oxygen probes.
- Total phosphorus (TP) and total nitrogen (TN) were measured using standard methods as previously described [74].
- Total organic carbon (TOC) concentrations were obtained by analysis on a Shimadzu TOC-L with sample changer ASI-L (Shimadzu Corporation, Japan).
- The concentrations of Fe(II) and Fe(III) were measured with the ferrozine colorimetric method [75].
- Gas concentrations were measured with a gas chromatograph (GC) (Agilent Technologies 7890A GC Systems) equipped with a flame ionization detector.
- Bacterial numbers were quantified on a flow cytometer (Cyflow Space, Partec, Görlitz, Germany) according to the protocol of Del Giorgio and colleagues [76].

5. Results and discussion

5.1 Paper I

In **paper I**, we designed and validated our targeted-metagenomics protocol for use with the newly acquired Illumina <MiSeq> sequencing instruments. At the time, several approaches using the Illumina technology for targeted-metagenomics had been published [77] [78] [79] [80] [81] [82] [83]. Nevertheless, we wanted to introduce our own protocol starting with PCR amplification of bacterial 16S rRNA genes, followed by paired-end Illumina sequencing and ending with bioinformatic analyses. We compared the reproducibility of results obtained and their correspondence with the results produced by the previous generation of sequencers used. We checked for potential biases and effects of different laboratories procedures as well as different downstream analysis algorithms.

In detail, our experimental Illumina amplicon sequencing design used barcoded primers flanking the V3-V4 segment of the 16S rRNA gene, a region commonly amplified in <pyrotag> experiments [84] [85] [86] [87]. To construct a standard Illumina paired-end library with an individual MID, 50 samples were amplified using a two-step PCR, mixed and then used as templates for MID coded library preparation. Usually 4-6 of these MID coded libraries were then simultaneously sequenced on an Illumina MiSeq.

Using a read length of twice 250 bp, the V3-V4 region of the rRNA gene presents an optimal target for sequencing [88] as it provides an adequate overlap of the forward and reverse paired-end reads. Moreover, assembling these reads increases the quality and confidence in the overlapping region [80] [82]. To reconstruct the complete nucleotide sequence we used the PANDASeq algorithm [89]. Next, as an exact clustering algorithm that computes the difference between every pair of sequences scales with the square of the number of input sequences cannot be used on a dataset of this size, we used UPARSE [90]. For every OTU obtained by UPARSE, the representative sequence of the cluster was used as a query against the quality checked SILVAMOD database using CREST [91].

Finally, a number of widely used statistical approaches were implemented in the analysis pipeline such as NMDS ordination plots, beta-dispersion, PERMANOVA, permutational ANOVA and the estimation of ecological diversity indices. The Bray-Curtis distances were calculated with the usual square transformation and Wisconsin standardization using rarefied datasets. In addition, tools to calculate the UniFrac distance [92] were implemented in the pipeline.

We concluded that the method worked well and was suited to the scientific questions we had in mind, as it can process large number of samples in parallel. The precisions of measured alpha and beta-diversity estimates were high and the effects of individual barcodes were negligible. Our comparative analyses suggest that Illumina and 454 data can be used in combination as long as the same PCR protocol and bioinformatic workflows are applied for describing patterns in taxonomic composition. This method and its updated versions have already been applied in multiple research projects such as the following: [93] [94] [95].

5.2 Paper II

In **paper II**, we made use of online databases and the information publicly available to add context to our own sequencing datasets. We built a tool titled `seqenv` that assigns linear combinations of environmental terms for individual sequences by performing homology searches on NCBI's "nt" database and parsing the results with a text miner that identifies occurrences of a set of controlled vocabulary. In this manner, we are able to quantify sequences and samples by a combination of terms describing the scope and niches of its microbes.

Briefly, `seqenv` automatically performs similarity searches of short sequences against the "nt" nucleotide database provided by NCBI and, for every hit, extracts the <isolation source> textual metadata field. After collecting all the isolation sources from all the search results, it runs a text mining algorithm to identify and parse words that are associated with the Environmental Ontology (EnvO) controlled vocabulary. This, in turn, allows us to determine both in which environments individual sequences or taxa have previously been observed and, by weighted summation of those results, to summarize complete samples.

We applied `seqenv` on a dataset of 18S rRNA OTUs retrieved from 48 sediment samples of a Black Sea core spanning the last 11'400 years [96]. Using a random forest classifier on the EnvO terms associated with each sample we were able to predict the past stages of the Black Sea. This classifier had an error rate of 12.5%. In figure 19, we show the relative frequency of the ten most important terms in this classifier across the samples, ordered by age and with the environmental stages indicated.

The oldest stage, ES4 or lacustrine interval (11.4–9.3 thousand years, ky, before present, B.P.) represents a stage where the Black Sea was disconnected from the Mediterranean Sea due to low sea levels. The sediment layers evidenced a strong signal connected to freshwater eukaryotes corroborating previous conjectures that, at that time, the Black Sea was akin to a freshwater lake. This phase ends with the initial marine inflow as, due to the end of the ice age 11'700 years ago, rising sea levels resulted in the connection of the Black Sea to the Mediterranean.

A period of increasing salinity (9.3-7.5 ky B.P.) followed, corresponding to the warm and moist mid-Holocene climatic optimum. As shown in figure 19, marine and brackish eukaryotes increased while freshwater eukaryotes decreased, supporting the accretion of salinity in the Black Sea. Further gains in salinity associated with the onset of the dry Subboreal initialized the establishment of modern environmental conditions (7.5 ky B.P-2.6 ky B.P.). The most recent sediment layers showed amplified freshwater signals coinciding with the freshening (2.6 ky B.P.-present) of the Black Sea, coinciding with the onset by the cool and wet Subatlantic climate and recent anthropogenic perturbations.

In addition, we applied `seqenv` to a dataset of archaeal amoA genes (ammonia oxidizing) from 45 British soils, originating from a broad range of pHs (min. 3.5, max. 8.7, median 6.2) [97]. The goal was to establish the biogeography of ammonium oxidizing soil archaea (AOA). Correlation analyses revealed a positive relationship between the diversity of habitats and the pH of the samples the organisms were found in (adjusted R-squared: 0.274, p-value: 3.85e-06). Another positive association was observed between sample pH and OTU diversity (adjusted R-squared: 0.131, p-value: 0.00922). To determine which environmental terms were most associated with the pH preference of the AOA OTUs, we again performed a random forest regression of pH preference against the weighted environmental terms. This model explained 34.6% of the variation in pH preference. Figure 17 gives the weights of the top ten most important terms as determined by the statistic "percentage increase in mean squared error of predictions", across OTUs ordered by their pH preference.

An early version of `seqenv` was already applied to bacterial communities of sediment cores retained from Swedish lakes [94]. This study revealed microbial community patterns coinciding with the history of their isolation from the Baltic Sea. Another early application of `seqenv` identified dispersed and dormant marine bacteria in freshwater lakes and the air of the Scandinavian mainland [98].

5.3 Paper III

In **paper III**, we used the two methods developed, as well as the expertise built, and applied it to the Danube river in the context of the second "Joint Danube Survey". We analyzed the bacterioplanktonic composition of 137 samples collected along the midstream of the 2'600 km from source to mouth. Our results showed that, overall, bacterial richness and evenness declined downriver and that the effect of tributaries was negligible in determining the microbial population. Finally, we compared these findings with previous theoretical ecology frameworks such as the river continuum concept.

Regarding bacterioplankton in large river networks, we propose, based on our own as well as other studies [99] [100] [101], that the highest diversity

exists in headwaters, and from thereon decreases towards river mouths. This, we argue, results from the inoculation of bacterioplankton by advection from surrounding environments (i.e. soil and groundwater) in the headwaters as supported by results provided by `seqenv`. In lotic environments, bacterioplankton is transported primarily passively. Facilitated by the large contact zone of small headwaters, expressed also by large surface-area-to-volume ratio, the surrounding environment (soil and groundwater) contribute allochthonous bacteria to the river community [99] [100] [101]. These source environments of inoculation harbour a much higher diversity than aquatic communities. These newly introduced allochthonous bacteria should be at least temporarily capable of proliferating in their new lotic environment, making them constitutive members of the community. Overall, this process of allochthonous input can be described by the so-called "mass-effect", where dispersal of organisms exceeds the rate of local extinction [102] [99] [100].

Flowing downriver, with increasing river width and decreasing "riparian influence", we propose that "species-sorting" progressively prevails over "mass effects" in shaping the bacterioplankton composition. This is supported by the increase of the core communities' relative abundance in both size fractions, as well as the rapidly decreasing number of first-time occurrences of OTUs from upstream to downstream. The associated progressive rise of competitive taxa is substantiated by the observed simultaneous decrease in evenness together with bacterial richness in both size fractions downriver, as seen in figure 22. Lastly, the decline of cell volumes along the Danube River [103] as well as a rise of typical freshwater bacteria (figure 24), representing small cells with oligotrophic lifestyles [104] [105], provides further evidence for the increasing importance of "species-sorting". The growing resemblance with lake communities downriver (online figures S5B and C) further corroborates the idea that lake and river bacterioplankton resemble each other.

Considered together, the findings of this paper and the existing literature [101] [106] [100] [107] suggest that the diversity of bacterioplankton decreases from headwaters to the river mouth due to the decreasing importance of the "riparian influence". This is consistent with the important role the "river continuum concept" assigns to the riparian zone as well as to the physical drivers such as river flow and wetted perimeter. The increase in relative abundance of typical freshwater taxa downstream and decreasing influence of soil and groundwater bacteria downstream suggests that the "river continuum concept" is a valid interpretive framework. The concept stipulates a continuous gradient of physical conditions that elicit a series of biological responses, resulting in consistent patterns of community structure and function along the river system.

5.4 Paper IV

In **paper IV**, we made use of even larger quantities of data and formed depth profiles by applying shotgun-metagenomic sequencing on a series of ice-covered lakes in the boreal landscape. To analyze the results, we employed yet a different series of bioinformatic tools to assemble, merge, cluster and annotate continuous pieces of DNA sequences into composite genomes.

A bioinformatic pipeline including the filtering of reads based on their PHRED quality scores using sickle [108] was developed. Using Ray [109] the pipeline then creates a number of assemblies with different k-mer sizes (51, 61, 71 and 81 bp) with resulting contigs being cut into 1'000 bp pieces. These pieces are reassembled with Newbler (454 Life Sciences, Roche Diagnostics). Coverage is computed by running bowtie [110] to map the reads back to the Newbler-produced assembly. Duplicates are removed by picard-tools and for computing coverage, bedtools [111] is used. Once contigs and scaffolds shorter than 1 kb are discarded, concoct [54] is run for binning contigs to population genomes. These bins, often referred to as metagenome-assembled genomes (MAGs), are then separated into low quality and high quality groups, based on results obtained from CheckM [112]. A completeness value of over 60% and a contamination metric below 10% is the criteria set for classifying a bin as <good> (high quality).

To provide functional annotations on predicted proteins from the assembly, Hmmsearch [113] was run on the PFAM-A database [114]. Coverage information and abundance of proteins allows us to construct estimated genome equivalents of individual PFAMs by considering 139 PFAMs predicted to occur as a single copy in all genomes [115]. Annotations of high-quality bins were performed using the "MicroScope" pipeline [116] [117] with automatic annotations assisted by manual curation as described in the integrated bioinformatics tools and the proposed annotation rules.

Applying our pipeline we revealed the partitioning of metabolic functions along the redox tower, as described in figure 31. We were able to reconstruct genomes with capabilities for photoferrotrophy and anaerobic methane oxidation, congruent with our geochemical observations. While sequences from well described branches of life, including certain methanotrophic and sulfur reducing *Proteobacteria* were present, most of the sequences were from less studied clades. These clades represent a large part of the phylogenetic tree of life, but are all in poorly documented areas, where scientists are only beginning to gather detailed information concerning the metabolic repertoire found in their genomes.

By summarizing chemical, taxonomic and genomic information, the results highlight the importance of lithotrophy and anoxygenic photosynthesis as metabolic processes occurring in these systems, in particular through their role in carbon fixation, and sulfur and iron cycling. The later are tightly linked to the degradation of allochthonous organic matter in these net-heterotrophic lakes, since

the availability of particular sulfur and iron compounds as electron acceptors dictates the degradation pathway of organic matter. Moreover, the ratio of methane to carbon dioxide emitted per organic carbon entering the lake depends on how organic carbon is transformed through the various processes in the lakes interior. A schematic overview of the complex metabolism along the redox tower in these lakes is shown in figure 31.

The distribution of the redox zones was unique to of each individual lake, and with the uniqueness of the associated microbial taxonomic and functional diversity described throws doubt on current efforts to predict the responses of lakes to environmental changes caused by global warming.

Based on our in-depth characterization of three ice-covered systems, we conclude that predicting the consequences of global environmental change, such as the effects following the shorting of ice cover periods, critically depends on the response of the complex microbial mediated metabolism along each individual redox tower. Our study provides a first important step towards understanding how the millions of boreal lakes currently function and may react in the future.

5.5 Paper V

In **paper V**, we set out to Austria and performed metagenomic sequencing on a series of shallow alkaline <soda lakes>. We obtained genetic material from microbes similar to those found in freshwater to polar alkaline and hypersaline environments depending on the prevalent environmental conditions occurring over the annual cycle. This complies with "the everything is everywhere, the environment selects" postulation by Baas Becking [29].

Applying a shotgun-metagenomic approach as in paper IV, we were able to reconstitute the composite genomes of the abundant microbial inhabitants and to describe the organisation of microbes and their adaptations to alkaliphilic and psychrophilic environmental conditions. These adaptations to alkaline and psychrophilic conditions are of immense interest as they may lead to a multitude of biotechnological applications. The microbial genomes recovered contained many functional traits. Most notably, we show that the microbes employ multiple mechanisms to cope with the alkaline conditions.

In high pH environments, many cells have an inverted pH gradient (i.e. acidic inside) which results in decreased proton motive force [118]. To compensate for this energetically unfavorable state, most alkaliphiles keep high negative values of membrane electrical potential [118] by employing a sodium-pumping NADH-CoQ reductase (NQR) [119]. Establishing complete or partial operons of the rnf complex (rnfABCDG) containing homologs to the NQR genes encoding the sodium-translocating NADH:quinone oxidoreductase [120] in multiple recovered genomes as shown in figure 38 was therefore straightforward. We also brought to light the presence of operons that encode for proteins

to establish sodium-motive force (SMF) including a Na^+-motive cytochrome c oxidase exporting sodium from the cell. In addition, several SMF consumers such as Na^+-motive transporters, symporters, and a Na^+-type flagellar motor further emphasize the importance SMF at high pH while H^+-motive homologues operate at neutral and low alkaline pH or under lowered buffering capacity.

Another challenge of high pH environments, in particular for primary producers, is carbon fixation. High pH is associated with a depletion of CO_2 with HCO_3 becoming a more dominant species [121]. Cyanobacteria have developed CO_2 concentrating mechanisms (CCMs) such as active CO_2 and HCO_3 uptake systems and an internal micro-compartment (carboxysome) where the CO_2 level is elevated around the active site of the RuBisCo. This process allows the RuBisCo to operate at near V_{max} speed and implies a much smaller investment of nitrogen in RuBisCo to achieve a particular rate of photosynthesis [122]. The active uptake systems for CO_2 and HCO_3, as outlined by [123] and which were identified in our genome bins, were (i) a putative Na^+-dependent HCO_3 transporter and (ii) a constitutive CO_2 uptake system, NDH-1 dehydrogenase complex. In addition, we identified genes encoding for carboxysomes in two genome bins, a sulfur oxidizer, related to *Thioalkalivibrio*, and a photosynthetic *Synechococcaceae*. Based on our genome bins, cyanobacteria share all of the carboxysomal proteins so far described from *Proteobacteria*, as shown previously [124].

Our investigation provides important data concerning the wealth of enzymes present potentially tuned towards alkaliphilic conditions. Further, these might prove useful for biotechnological purposes, such as enzymes that can hydrolyze lignocellulose and other biopolymers (e.g. proteins, cellulose, xylan, and chitin). To conclude, this first explorative study of the Austrian soda lakes is a spur for future research in these easily accessible and extreme environments. This could include functional metagenomics studies and attempts to isolate some double-extremophilic prokaryotes.

6. Conclusion and perspectives

Microbes determine the health of plants and animals over the globe and dictate the planetary biogeochemical cycles. Yet, our knowledge on their make up, the rules determining their distribution in the environment and how their complex interactions determine ecosystem features, are poorly described and understood. In this thesis, several bioinformatic tools were developed in order to process and analyze metagenomic sequence data. First, a tool for targeted-metagenomics was programmed to assess the diversity of bacteria, archaea and eukaryotes at high throughput permitting the study of biogeographic patterns in great detail. Second, a software package providing annotations based on environmental ontology terms was created. Its applications are in microbial source tracking, and studies of paleontology and biogeography. The studies presented confirm the view that the dispersal limitations of microbes are more or less non-existent because environmental properties dictate their distribution. The results also attest to the fact that microbes can be used to trace the origin and history of the sampled communities.

Third, a shotgun-metagenomics analysis pipeline was set up enabling the reconstruction of near complete genomes from uncultured microbes. This tool was applied to the investigation of the biochemical processes occurring in a selection of systems representative of the tens of millions of lakes and ponds of the boreal landscape. The pipeline was also used in an another study concerning alkaliphilic lakes and revealed energy acquisition and carbon fixation strategies under alkaline conditions.

These studies are perfect illustrations of the "reverse ecology" approach. In reverse ecology, the scientist starts with the genetic information obtained from an organism or an environmental sample and attempts to predict the identities, functions, or phenotypes of the community or individual cells inhabiting the environment under study. In contrast with *forward ecology*, where one would commence with the observation of a phenotypic property. In particular, we show that metagenomics approches produce data very suitable for ecology and allow to perform "high-throughput ecology" by providing a bridge between theoretical and analytical tools. By assessing their genomic content, we were able to measure community patterns and uncover some of the putative roles of microbes in elemental cycles. In addition, we could identify novel pathways and obtain the partitioning of metabolic processes in natural environments, factors which ultimately determining total ecosystem functioning.

One of the most obvious improvements in the bioinformatic processing of metagenomic data is the need for better assemblers. The assembly step which

starts by taking all the raw reads and produces a distribution of contigs, is crucial in correctly retrieving the genomes of the different microbes and doesn't always discriminate correctly between two populations, creating composite genomes. One of the issues is that none of the solutions available use the whole panoply of information available to make decisions, always trying to solve solely one sub-problem at a time. For instance, the <joining> and <binning> are always separate and would need to be merged within one comprehensive software to create a more holistic process. Currently, most of the joining algorithms (also simply called assemblers) only loads individual sequence reads in memory to produce contigs, but ignore which sequence came from which sample. The binner algorithm, conversely, only loads the contigs and a matrix indicating the coverage obtained in each sample in memory. What we need is one program that would load all available information in memory to make use of more complete and accurate heuristics to determine the correct path in the De Bruijn graph. This would include, sequences, their associated quality, the provenance of each sequence (to use differential abundance strategies), an analysis of the sequence composition (e.g. tetranucleotide frequency) and, additionally, information retrieved from known completed genomes. The direct visualization and manipulation by the scientist of this process is also a goal that is sorely missed by current algorithm developers.

This thesis revolved around the use metagenomics techniques for microbial ecology, but several other promising techniques are being developed. Multiple different approaches proposed over a decade ago [1] exist, with many still in their infancy. The three main categories of experimental systems are outlined in figure 5. Most notably, single-cell methods which avoid the problem of co-assembly entirely as each sequence recorded necessary is contained in the same individual either as DNA, RNA or AAs. This type of technique can give us precise insights into the variation among individuals which is a central feature of cellular life and a major theme of biology. For plant and animal populations, genotypic and phenotypic variations at the level of individuals is well known to be associated with social (intra-population) and ecological (inter-population) interactions. These interactions have been shown to create frequency-dependent selective pressures that can lead to evolutionary changes in response to environmental conditions. However, in the case of prokaryotes, we are only beginning to learn how diversity among closely related genotypes mediate ecological and evolutionary processes.

Yet another approach is that of enrichment cultures and mixed cultures which have been proposed as testbeds for ecological and evolutionary principles since they are less complex than natural systems but maintain many of the population's interactions intact [244]. Once again, these simplified communities can then be subjected to various analyses targeting either their genome, their transcripts or their proteins, as well as other chemical and molecular protocols, all with large potential for gaining insights into ecological interactions and co-evolution. However, to get the experimental design right, the settings of the

systems of interest have to move into the center of the microbial ecologist's mind as we learn to master the technical challenges.

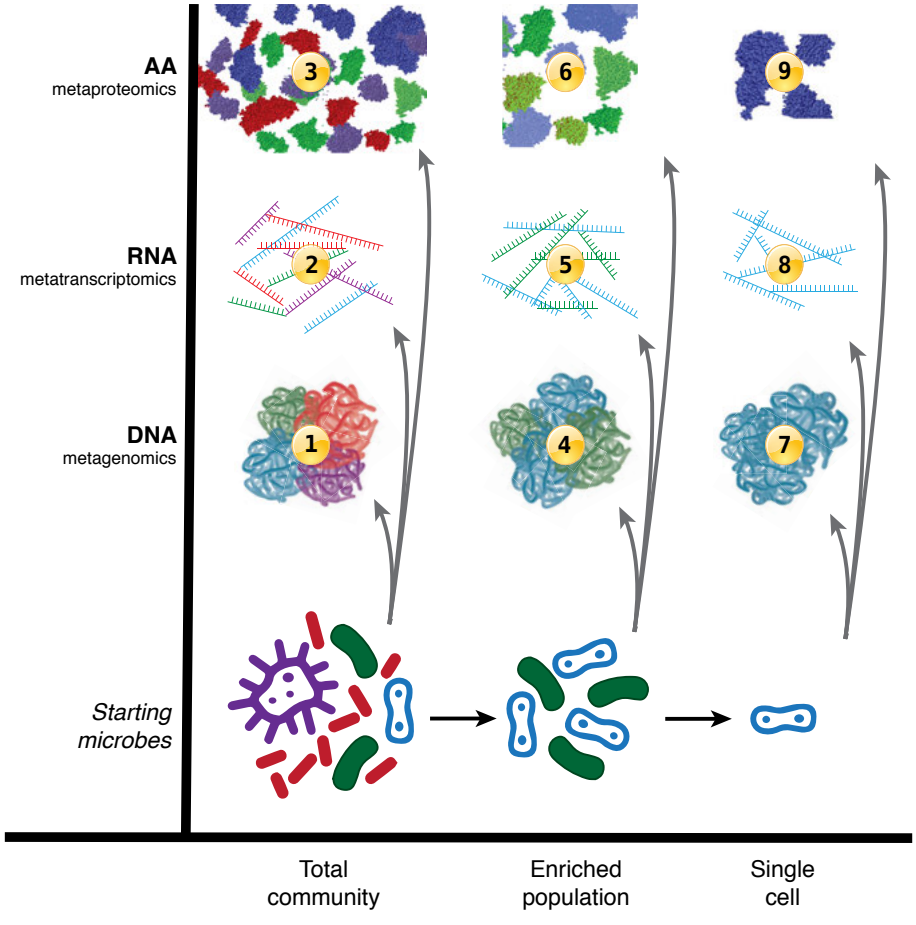

Figure 5. Different methods for microbial ecology. Adapted from [1].

A microbe is the closest thing we have to a programmable, self-reproducing and self-sustainable nanorobot today. Yet, we are far from the level of scientific and technical knowledge that would allow us to truly engineer them and unlock their potential. In fact, we barely understand what causes them to assemble and behave as they do. By better comprehending and controlling microbes we could make progress by leaps and bounds in areas such as medicine and bioremediation, as well as in atmospherical manipulation as microbes can catalyze useful reactions on planetary scales. Some even argue that microbes are the best tool we currently have to attempt exoplanet terraforming.

7. Summary in Swedish

Mikrobiell ekologi syftar till att förstå och förutsäga förhållandena mellan mikroskopiska organismer samt deras interaktion med den omgivande miljön. Med hänsyn till den minimala storleken och vår oförmåga att manipulera och kontrollera de flesta mikroorganismer, finns många metodologiska begränsningar inom vetenskapsområdet. Till exempel har klassiskt skolade mikrobiologer under det senaste århundradet i stort sett misslyckats med att isolera och renodla majoriteten av de mikrobiella arter som finns i naturen. Oavsett hur vi ändrar rådande miljöförhållanden så verkar mindre än 0,1 % av naturens mikroorganismer kunna föröka sig och bilda synliga kolonier på näringsagar.

Numera övervinns vissa av dessa metodologiska begränsningar genom användandet av molekylära metoder. Dessa är numera de mest använda metoderna inom mikrobiell ekologi. Särskilt populär är tillvägagångssättet att med höggradig parallellisering utföra genetisk sekvensering av organismernas arvsmassa. Dessa resurskrävande laboratorieverktyg gör det möjligt att läsa det universella genetiska alfabetet som består av de fyra baserna, A, T, G och C. För att till fullo beskriva och tolka de stora datamängder som erhålls behövs väldokumenterade och kraftfulla bioinformatiska verktyg som gör det möjligt för forskare att beskriva biologiska system med avseende på bland annat mångfald och funktion.

I denna avhandling har flera bioinformatiska verktyg utvecklats och validerats för att bearbeta och analysera sekvensinformation från komplexa mikrobiella samhällen, så kallad metagenomiska data. Därefter har dessa verktyg använts för att förbättra vår förståelse av mikrobiell biogeografi och systembiologi. Det första verktyget som beskrivs i avhandlingen är inriktad på biogeografiska studier. För detta ändamål inriktat sekvensanalysen på en specifik markörgen som återfinns i alla levande organismer då den är nödvändig för cellens grundläggande funktion. Denna metod-ansats bär det passande namnet “riktad metagenomik”. Genom att begränsa sekvensering till endast en liten men jämförbar del av arvsmassan får man en möjlighet att kvalitativt mäta egenskaper hos det mikrobiella samhället såsom mångfald, jämnhet och artrikedom. Genom att jämföra med databaser kan man även beskriva mikroorganismerna i taxonomiska termer. Metodiken är väl lämpad för att identifiera och påvisa tidigare beskrivna bakterier eller arkéer. Som en del av detta avhandlingsarbete utvecklades ett datahanterings och tolkningsverktyg för ovan beskrivna riktade metagenomik-data, som automatiserar kvalitetsfiltrerar, sammanfogar och taxonomiskt bedömer den mångfald av bakterier, arkéer och eukaryoter och möjliggör vidare studier av biogeografiska mönster i mer detalj.

Därefter utvecklades en andra mjukvara som kopplar sekvensinformation till den miljö som organismerna återfinns i. Detta med syfte att utnyttja det överflöd av existerande information som finns tillgänglig i offentliga databaser. Verktyget tillämpades därefter i studier som innefattade källspårning, paleontologi, och biogeografi. Båda dessa verktyg har redan bidragit till att utvidga vår förståelse av mikrobiell mångfald i inlandsvatten och har även bidragit till vår uppfattning om mikrobiell biogeografi i rinnande vatten. Datahanteringssystemen har använts för att analysera prover från flera miljöer, såsom alkaliska sodasjöar och gamla sedimentkärnor. Dessa studier bekräftade att mikroorganismernas spridning har få begränsningar varför miljöegenskaper dikterar deras fördelning i miljön. Mina resultat visar även att vilande mikrober kan användas för att rekonstruera ett tidigare livaktigt mikrobiellt samhälle.

Den andra analys-strategin som utvecklades och användes är inte behäftad med någon begränsning till en specifik region av genomet utan utgår istället från bred sekvensering av allt genetiskt material i det analyserade provet på ett i stora drag slumpmässigt sätt. Denna metodik går under beteckningen "shotgun-metagenomik" och ger oss möjligheten att även beskriva metabola funktioner och andra cellulära egenskaper i organismerna. Alla reaktioner som en cell utför katalyseras av ett eller flera proteiner. De kodande regionerna i arvsmassan fungerar som ritningar för tillverkning av dessa proteiner, så med hjälp av denna typ av sekvensinformation kan man i kombination med rätt databaser och algoritmer förutsäga vilken metabolism som en viss av mikrobiell population drivs av samt räkna ut hur de bidrar till omsättningen av olika grundämnen och energi i ekosystemet där de förekommer. Det är i detta sammanhang viktigt att beakta mikroorganismernas centrala betydelse för vår planets biogeokemi, inte minst i denna tid när vi bevittnar snabba och dramatiska miljö- och klimatförändringar.

I arbetet med att utveckla bioinformatiska verktyg för denna typ av "shotgun-metagenomik" data har alla nödvändiga steg i analysprocessen, från bearbetning av rå sekvensdata till funktionell annotering och rekonstruktion av prokaryota genom, kopplats samman till ett väl fungerande arbetsflöde. Genom att tillämpa denna typ av verktyg kunde biokemiska processer i ett urval av akvatiska system representativa för de tiotals miljoner sjöar och dammar som finns i det boreala landskapet rekonstrueras. Detta avslöjade att många rikligt förekommande och hittills obeskrivna prokaryoter utför kritiskt viktiga biogeokemiska funktioner i dessa akvatiska ekosystem. Vi kunde även påvisa förekomsten av organismer med kapacitet för ljusdriven metabolism kopplat till järnoxidation och anaerob oxidation av metan. Detta är egenskaper som inte tidigare upptäckts i dessa system. I en annan studie visade vi att mikroorganismer svarar på alkaliska betingelser genom att justera sin anförskaffning av energi och strategier för att fixera kol för tillväxt. Sammanfattningsvis kan genomiska analyser vara ett mycket kraftfullt verktyg för att påvisa en potential för olika biogeokemiska processer i naturen, inte minst då möjligheten finns att detektera både komplexa och oväntade processer i naturliga miljöer.

Vi kallar detta för “omvänd ekologi” eftersom vårt angreppssätt beskriver organismer eller processer genom en bottom-up-strategi där man utgår från den genetiska informationen för att dra slutsatser om vår miljö eller de organismer som lever där. Idén till detta synsätt kan konkretiseras genom att reflektera kring de studier som Darwin genomförde på Galapagosöarna. Hur skulle dessa experiment ha sett ut idag med tillgång till modern DNA-sekvenseringsteknik? Vad Darwin gjorde för över hundra år sedan var att observera morfologin hos olika fink-arter på Galapagosöarna. Under detta arbete noterade han i synnerhet de olika typer av näbbar hos dessa fåglar. Utvecklingen av de olika morfologierna förklarades av tillgång till olika typer av födokällor och konkurrens. Detta var en fenotypisk observation som först långt senare kopplades till dess underliggande genetiska orsak, nämligen den uppsättning av gener som ligger till grund för den morfologiska utvecklingen av finkar och egenskaperna hos deras näbb. Detta är ett exempel på klassisk ekologi där man kopplar egenskaper i en viss miljö till variation i fenotypiska egenskaper och där forskare senare har kunnat arbeta sig ner till den genetiska bakgrunden till detta. I omvänd ekologi börjar forskningen istället med den genetiska information som erhållits från en organism eller ett miljöprov för att sedan försöka förutsäga deras identitet, funktion eller fenotyp. De teknologiska och beräkningvetenskapliga framsteg som vi just nu ser öppnar dörren för denna forskningsmetodik.

8. Acknowledgements

This is the part where you say thanks to all the people you met during the four years of your thesis. That would be a long list, I'll try to keep it short and meaningful.

Thanks to Mercè for being a cool office roommate along with Jürg upon my arrival. I enjoyed the swimming competitions! Then replaced by the equally interesting Martin. We had some fun badminton games!

Thanks to Sari for orchestrating and being instrumental in our sampling trip to Jämtland. Couldn't have made it without you.

Thanks to Moritz for our bioinformatics discussion and your advice.

Thanks to Sainur and Pilar for their invaluable work in the laboratory. I could not have processed all those samples and prepared them for sequencing.

Thanks to my Austrian collaborators Alexander K, Andreas and Domenico for providing samples and datasets used in this thesis.

Thanks to Stefan for commenting and suggesting revisions to this thesis.

Thanks to Valerie and Alina, the funest girls in the departement for our many after-works!

Thanks to all the other people in the Limnology departement and EBC. Was nice hanging out.

Thanks to Ludo, my friend and Swiss connection in Sweden.

Thanks to Prune, the coolest DJ Uppsala will ever have!

Thanks to Christian, my best roommate to date and good Swedish friend in my time aboard.

Thanks to Caro, I won't forget you.

A very big thanks to Chris Quince who provided essential ressources for the completion of the thesis. In both enlightening discussions and materials. What is a bioinformatician without a working and performant computer node?

Thanks also to the UPPMAX facility for granting access to their semi-reliable computer cluster all along my studies.

A deep gratitude to Frank and Anna for their support. It's meant a lot to me.

And of course, to Alex! Without whom I wouldn't be here and nothing would have taken place! Cheers to the best supervisor, truly a diamond amongst the stones. Needless to say, I'm looking forward to continue working with him at Envonautics Ltd.

Finally, the most important, thanks to my mother Anne!

It is common to add a famous citation or quote here. One of my favorites and befitting the "Ph" part of the abbreviation "PhD":

> The philosopher is like a man fasting in the midst of universal intoxication. He alone perceives the illusion of which all creatures are the willing playthings; he is less duped than his neighbor by his own nature. He judges more sanely, he sees things as they are. It is in this that his liberty consists - in the ability to see clearly and soberly, in the power of mental record.
> – Henri Frédéric Amiel of Geneva (1821-1881)

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1388

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-297613

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2016