Bielefeld University

Faculty of Technology
Center for Biotechnology (CeBiTec)

FROM SINGLE GENOMES TO NATURAL MICROBIAL COMMUNITIES: NOVEL METHODS FOR THE HIGH-THROUGHPUT ANALYSIS OF GENOMIC SEQUENCES

By

Lutz Krause

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR RERUM NATURALIUM

AT

FACULTY OF TECHNOLOGY

BIELEFELD UNIVERSITY

BIELEFELD, GERMANY

OCTOBER 2007

Lutz Krause
Teutoburgerstr. 47
33604 Bielefeld
lkrause@cebitec.uni-bielefeld.de

Supervisors: Prof. Dr. Jens Stoye
Prof. Dr. Alfred Pühler

Summary

Advances in sequencing technologies provide the opportunity to rapidly produce vast genomic data sets at a low price. In particular, the recently developed, ultra-fast 454 pyrosequencing has dramatically reduced the cost and time requirements per sequenced base pair. Furthermore, novel methods have been developed that enable sequencing of the 99% of microbes that are difficult to access with conventional, culture-dependent approaches. These culture-independent methods launch the exciting field of metagenomics – the study of the collective genomes (*metagenomes*) of free-living microbial communities.

In light of the immense data sets produced, sequence analysis is still a contemporary, ongoing challenge in computational biology. Additionally, new demands arise from the short length of sequence reads produced by ultra-fast sequencing techniques. In the field of whole-genome research, accurate methods are required for identifying and functionally characterizing the gene content of organisms, thus reducing the required manual effort while at the same time producing high-quality annotations. On the other hand, metagenomes are nowadays routinely sequenced, and an increasing number of metagenomic projects is expected in the near future (Pennisi, 2007), but their computational analysis is still in its infancy.

In the presented thesis, state-of-the-art machine learning techniques as well as algorithmic and statistical methods are employed for the high-throughput analysis and characterization of large genomic data sets. First, the gene finding software GISMO was developed, which combines the search for protein domains using profile hidden Markov models (pHMMs) with a sequence composition-based classification using a Support Vector Machine. This combined strategy is able to unveil almost the complete gene content of prokaryotic genomes in a fully automated manner. GISMO has already been extensively employed in the international effort to 'Annotate a Thousand Genomes' as well as in various genome annotation and re-annotation projects.

Furthermore, a novel gene finding algorithm for metagenomic data sets was developed. It is robust for most problems encountered when predicting genes in metagenomes, including short sequence length and low sequence quality. Thereby, the algorithm allows to hunt for novel, unknown genes carried by organisms that cannot be sequenced using conventional, culture dependent techniques.

Finally, methods were devised for characterizing short-read metagenomes obtained by pyrosequencing. Following the pHMM-based identification of gene fragments, the latter are categorized into functional groups. Additionally, the source organisms (taxonomic origins) of gene fragments are predicted. The resulting genetic and taxonomic profiles can in turn be used to unveil important trends in the gene content, metabolism, and species composition of the underlying microbial communities.

Contents

CHAPTER 1

Preface

Microorganisms are ubiquitous and essential components of all ecosystems on Earth and comprise more than half of its biomass (Fraser *et al.*, 2000a). They largely contribute to forming the geochemical composition of Earth's biosphere, buffer environmental changes, and hold a wealth of new functions for natural product synthesis. Exploring the microbial universe could reveal enzymes that hold the key for cleaning up pollution or for the industrial production of biofuel, food, and pharmaceuticals. Estimates suggest that by 2010 between 10% and 20% of all chemicals sold could be produced using biotechnological applications (Lorenz and Eck, 2005). Despite the importance and predominance of microbes, we have just started to reveal their incredible diversity, the full range of biochemical processes they are capable of, and their response to environmental changes. For example, the estimated number of at least 10 million bacterial species are in striking contrast to the few thousand that have been formally described (Gould, 1996).

A new era of exploring microbial diversity was initiated in 1977 when Carl Woese and colleagues assessed the evolutionary relationships of organisms by studying their 16sRNA and 18sRNA (Woese and Fox, 1977). Comparative sequence analysis of RNA genes revealed that cellular life can be divided into three domains, one eukaryotic (Eukaryote) and two prokaryotic (Bacteria and Archaea). This approach complements the traditional classification of microbes based on morphological characteristics, which does not provide a valid taxonomic grouping (Eisen, 2007). Microbes evolve so rapidly that two close relatives may have very different appearances, while distantly related species may look similar. Since the pioneering work of Norman Pace and colleagues in the mid 1980s, our knowledge about the diversity of microbes has largely been increased by directly isolating RNA genes from the environment, followed by their phylogenetic characterization (Pace *et al.*, 1985; Hugenholtz *et al.*, 1998; Hugenholtz, 2002).

The genomic revolution began between 1977 and 1995; first, Sanger *et al.* (1977a) sequenced a virus infecting *Escherichia coli*; then, in 1995, Fleischmann *et al.* published the first complete genome sequence of a free-living organism, *Haemophilus influenzae*. Ever since, many advances have been accomplished both in sequencing techniques and molecular methodologies, making microbiology one of the fastest progressing sciences. Nowadays, after cultivating microbes in the lab, their entire genomes are routinely sequenced in high throughput (Fraser *et al.*, 2000b). Studies of microbial genomes have tremendously increased our knowledge about the physiology, evolution, and biochemical processes of organisms (Nierman *et al.*, 2000; Fraser *et al.*, 2002). Results have given striking insights into disease causing mechanisms of microbes and have provided novel approaches in treating microbial infections and designing antimicrobial agents and vaccines. However, estimates suggest that only 1% of microbes can be cultured using standard techniques (Hugenholtz *et al.*, 1998; Hugenholtz, 2002). Another big step forward was the development of cultivation-independent methods to directly isolate microbial DNA from the environment. Using these methods, it has been possible to sequence microbes that were previously considered to be inaccessible, such as obligate pathogens (Fraser *et al.*, 1998) or symbionts (Shigenobu *et al.*, 2000), which cannot survive outside their hosts, or even ancient DNA from long-extinct species like Neanderthal (Green *et al.*, 2006; Noonan *et al.*, 2006) or Mammoth (Poinar *et al.*, 2006). However, whole-genome studies of microbes have been highly biased towards the four phyla proteobacteria, firmicutes, actinobacteria, and bacteroidetes, which therefore dominate our current knowledge about microbiology (Hugenholtz, 2002).

At present, we are just awakening to a new revolution, the sequence based study of microbial communities in their natural habitats, called community genomics, ecogenomics, or metagenomics (Riesenfeld *et al.*, 2004; Schloss and Handelsman, 2005; Tringe and Rubin, 2005). Over the past decade, microbial genome research has mainly focused on studying individual species that could be cultivated. However, in nature microbes are embedded in dynamic communities of multiple coexisting species with complex interactions. The natural life of microbes can therefore only be revealed by studying communities in their natural environment instead of single species cultivated in the lab. In metagenomics, the collective genomes of natural microbial communities are directly isolated from the environment and subsequently analyzed. These approaches give the opportunity to investigate the potential metabolism and species composition of microbial communities, while at the same time giving access to the genome sequence of the 99% of microbes that are difficult to cultivate. Moreover, metagenomic approaches provide a comprehensive view of the interactions, higher-level evolution, dynamics, and response to environmental perturbations of complex microbial communities found in nature.

Nowadays, molecular biologists are capable of producing large amounts of genomic sequence data within hours, raising the demand for novel computer programs to handle them. The goal of this thesis is to develop computational methods that aid researchers in analyzing, characterizing, and interpreting genomic and metagenomic data sets, thus helping to increase our knowledge on the hidden world of microbes. The three main goals are:

First: In light of the pace at which genomes are sequenced today, more accurate gene finding programs are required to further reduce the manual effort involved in genome research. Motivated by this demand, the first goal is to develop accurate methods for identifying the gene content of prokaryotic genomes in a fully automated manner.

Second: Owing to their complexity, many existing computational sequence analysis tools cannot be applied to metagenomic data sets. Hence, the second goal is to develop gene finding methods for metagenomes that can be used to hunt for novel, unknown genes that are not harbored by any of the cultivated organism.

Third: The recently developed, ultra-fast 454 pyrosequencing has significantly dropped the cost and time requirements of sequencing. However, because of the short average length of 100 bp of the reads produced, their computational analysis is a major challenge. The third goal of this thesis is to devise methods for characterizing short-read metagenomes obtained by pyrosequencing. For a given metagenome, the developed methods should be able to decipher the gene content and species composition of the underlying microbial community.

Overview

In the following, the structure of the presented thesis is outlined. Chapter 2 introduces the background and basic terminology required to describe the results of this work. First, genome research of single prokaryotes and entire microbial communities are presented, followed by the introduction of computational methods used in the analysis of molecular sequences. Next, the Support Vector Machine classification technique is introduced, which in the context of this work is employed to identify protein encoding genes in complete prokaryotic genomes. Finally, measures used to evaluate the accuracy of a classifier as well as measures for characterizing the biodiversity of microbial communities are described. Chapter 3 provides an overview of the contributions accomplished by the author in the analysis of whole prokaryotic genomes and microbial metagenomes. For this purpose, six selected scientific articles are summarized. Chapter 4 gives a synopsis on the new findings that were gained during this work when analyzing and characterizing genomic data sets. Chapter 5 considers future directions. In the last Chapter, the six scientific manuscripts included in this thesis are presented in the layout of the journals in which they were published.

CHAPTER 2

Background

2.1 Whole-genome research of prokaryotes

Whole-genome projects aim to gain new insights into the biochemical and physiological processes carried out by organisms, based on studying the information encoded in the linear nucleotide sequence of their DNA (Fraser *et al.*, 2000b; Nierman *et al.*, 2000). Nowadays, the nucleotide order of entire microbial genomes is routinely deciphered using high-throughput *shotgun sequencing*: After growing the target organisms in a culture, their genomes are randomly sheared into short fragments. These fragments are then inserted (*cloned*) into so called vectors (usually small DNA rings), and resulting clones are amplified and purified (Fleischmann *et al.*, 1995; Fraser and Fleischmann, 1997). Sequence reads of about 1000 bp are obtained from randomly selected fragments using *Sanger sequencing* (Sanger *et al.*, 1977b). Finally, the original sequenced genome is *reassembled* (or simply *assembled*) with powerful computer programs that link overlapping reads into longer, continuous stretches of DNA, called *contigs* (Sterky and Lundeberg, 2000).

In contrast to the conventional approach outlined in the preceding paragraph, the recently published, massively parallel *454 pyrosequencing* technique (Margulies *et al.*, 2005) allows to directly sequence shotgun fragments without a prior cloning step. While this technique at the same time significantly drops the cost and time requirements per base pair, its central weakness lies in the short length of sequence reads produced, with averaging length of 100 bp.

After a genome is reassembled, computational methods are employed to obtain a preliminary description (*annotation*) of its sequence. Functional elements, including genes and regulatory elements, are identified and characterized. Metabolic pathways implemented by the target organism are reconstructed and events that shaped a genome during course of evolution are revealed, such as gene loss, transfer of genetic material, or rearrangements. These analyses are usually automated and constitute the first step in whole-genome studies, followed by experts manually polishing the annotations.

2.2 Metagenomics

Advances in sequencing methods have recently dramatically changed the way microbial ecosystems are studied, ushering into the emerging field of *metagenomics*, the sequence-based study of naturally occurring microbes. In metagenomics, the collective genome (*metagenome* or *microbiome*) of coexisting microbes – called microbial *communities* (Fauth *et al.*, 1996) – is randomly sampled from the environment and subsequently sequenced (Breitbart *et al.*, 2002; Tyson *et al.*, 2004; Venter *et al.*, 2004; Riesenfeld *et al.*, 2004; Schloss and Handelsman, 2005). Analogous to whole genomes, the most widely used strategy for sequencing these *environmental DNA samples* is the shotgun approach, resulting in a mixture of short genomic fragments of unknown origin (Eisen, 2007). While at present shotgun fragments are predominantly sequenced using the Sanger technique, the number of metagenome projects using 454 pyrosequencing is rapidly increasing (e.g. Edwards *et al.*, 2006; Turnbaugh *et al.*, 2006; Angly *et al.*, 2006). The 454 technique has significantly dropped the time and cost constraints of environmental DNA sequencing while at the same time avoiding the potential bias that the cloning procedure might introduce (Margulies *et al.*, 2005; Edwards *et al.*, 2006; Turnbaugh *et al.*, 2006).

By directly accessing the collective genome of co-occurring microbes, metagenomics has the potential to give a comprehensive view of the genetic diversity, species composition, evolution, and interactions with the environment of natural microbial communities (e.g. Béjà *et al.*, 2000; Furrie, 2006; Martin *et al.*, 2006; Whitham *et al.*, 2006; Gill *et al.*, 2006; Hansen *et al.*, 2007). Moreover, considering that the vast majority of microbes resists cultivation using conventional methods (Hugenholtz, 2002; Rappé and Giovannoni, 2003), metagenomic approaches highly enlarge our window into the microbial universe.

Studies based on metagenomics have greatly increased our knowledge about the diversity of microbes (e.g. Venter *et al.*, 2004; Bohannon, 2007; Rusch *et al.*, 2007); revealed changes in the gut microbiome of obese mice (Turnbaugh *et al.*, 2006); and expanded our understanding about the evolution of species interactions (Hansen *et al.*, 2007). Two of the most impressive metagenomic studies discovered photosynthetic marine gammaproteobacteria (aerobic anoxygenic phototrophs, AAnPs) (Béjà *et al.*, 2002) and bacterial rhodopsin (proteorhodopsin) (Béjà *et al.*, 2000). These findings challenge our view of the nature and prevalence of light-utilization strategies of microbes living

in ocean surface waters (DeLong, 2005). Further studies have provided glimpses into the gene reservoir of uncultured microbial communities from various environments, including the ocean (DeLong, 2005; Rusch *et al.*, 2007), whale falls (Tringe *et al.*, 2005), sewage sludge (Martin *et al.*, 2006), or human guts (Gill *et al.*, 2006). Forest Rohwer and colleagues have also used sequencing techniques to explore the composition and gene content of viral assemblages in their natural environments (Breitbart *et al.*, 2002, 2003, 2004a,b; Edwards and Rohwer, 2005; Angly *et al.*, 2006).

Characterizing the species composition and genetic potential of microbial samples, in other words, revealing which microbes live in an ecosystem and what they are doing, are among the most essential questions in metagenomics and challenging tasks in computational biology. For the first task, the source organisms or *taxonomic origins* of environmental genomic fragments are inferred. In biology, organisms are systematically classified into categories sharing particular characteristics. These categories, called *taxonomic groups* or simply *taxa*, are organized in a hierarchical manner into the ranks of superkingdom, phylum, class order, genus, and species. Superkingdom is the highest-level grouping of cellular live. By assigning a taxonomic origin to each environmental genomic fragment, a community specific taxonomic profile is derived. These profiles portray the species composition of the underlying communities on each taxonomic rank, in other words characterizes their *taxonomic composition*.

For the task of characterizing the genetic potential of microbial communities, protein encoding sequences (*coding sequences, CDSs*) are identified in the genomic fragments of a sample and functionally described. In case genomic fragments are directly analyzed without a prior assembly step, detected coding sequences are called *environmental gene tags* (*EGTs*) – snippets of genes that potentially encode a protein adapted to the environment (Tringe *et al.*, 2005). After assigning a cellular function to each EGT, the resulting profiles can be used for quantitative gene content analysis in order to reveal habitat-specific genetic fingerprints. Genes that are important for survival in a habitat or adaptation to a niche, will occur in the genomes of many organisms sharing the same environment (Tringe and Rubin, 2005). By comparing the EGT profiles of unassembled metagenomes, these community and environment-specific genetic traits as well as trends in the metabolic activities and functional roles of each microbial community can be unveiled.

While environmental DNA is sequenced at a high rate and an increasing number of metagenome projects can be expected in the near future (Pennisi, 2007), the computational analysis of genomic fragments from a heterogeneous mix of species is still loaded with problems. For example, the short length of the sequence reads obtained combined with an inherent intra-species genetic heterogeneity (Spencer *et al.*, 2003; Thompson *et al.*, 2005), inter-species gene conservation, and variable species composition, all these together entail a fundamental computational challenge for characterizing the genetic and taxonomic composition of the underlying communities or for assembling each genome into longer contigs. Additionally, the main obstacle for assembling complex

communities is simply an insufficient sequence coverage. For example, half of the 7.7 million DNA sequence fragments that were collected by Venter and colleagues on their sailing cruise from Halifax to the Galapagos islands were so different that they could not be linked at all (Rusch *et al.*, 2007).

Despite the technical difficulties, for simple communities, metagenomes obtained by Sanger sequencing are successfully assembled into longer contigs if sufficient coverage is available (e.g. Tyson *et al.*, 2004). While from species that are over-represented in a sample long stretches of genomic sequences can frequently be assembled, for under-represented species only short contigs with low coverage are obtained. One problem related to low coverage is that contigs are even more prone to contain sequencing errors.

2.3 Analyzing genomic sequences

This section introduces computational sequence analysis methods that are employed in the context of the presented thesis. The described methods are used for a) the functional characterization of genes based on the detection of sequence similarities, b) the automated identification of protein encoding genes, c) reconstructing the evolutionary relationships (*phylogenies*) of molecular sequences, and d) predicting the source organisms (*taxonomic origins*) of environmental DNA fragments. Considering that genomes can be represented as linear sequences of nucleotides and proteins as linear sequences of amino acids, their computational analysis can be regarded as the analysis of sequences of characters over the alphabet of four nucleotides {A,C,G,T} and over the alphabet of all 20 amino acids. Each nucleotide and amino acid is represented by a single character: adenine (A), cytosine (C), guanine (G), thymine (T) for nucleotides or alanine (A),. . ., tyrosine (Y) for amino acids.

2.3.1 Sequence comparison methods

Commonly, the first step to infer the cellular role of new proteins involves a search for sequence similarities in a database of functionally characterized amino acid sequences. The similarities identified enable to transfer information about the function and folding from related proteins. The broad success of similarity based methods in computational biology can be explained by the phenomena that during evolution most new genes are not invented *de-novo*, but arise by duplication, rearrangement, and mutation events of existing genes (Chothia *et al.*, 2003). Related proteins stemming from the same ancestor are called *homologs* (Fitch, 2000).

To compare sequences, usually an *alignment* is computed, such that similar characters are placed in the same column, in respect to an internal scoring scheme (Durbin *et al.*, 1998). An alignment computed between two sequences is called *pairwise alignment*, otherwise *multiple alignment* when more then two sequences are compared. Various pairwise alignment methods have been implemented that employ a heuristic

to speed the search up, with perhaps the most prominent one being BLAST (Altschul *et al.*, 1997). For a given query sequence BLAST can, very time-efficiently, screen huge databases for similarities. Database sequences similar to the query are called *matching sequences* or simply *hits*.

In the following, profile methods and the Pfam protein families database are introduced. While the former are used to summarize the information contained in a multiple alignment, the latter makes extensive use of such profiles and in the context of this work is used for the detection of conserved coding sequences.

Profile methods

Profile methods are used to model multiple alignments of previously identified members of a protein family. The constructed models can in turn be employed to search for new family members. Among the most widely used profile methods in computational biology are position-specific scoring matrices and profile hidden Markov models (pHMMs) (Durbin *et al.*, 1998). Position-specific scoring matrices represent the likelihood of seeing each amino acid in each column of the underlying alignment. Hence, they represent which amino acids are allowed at any position, which positions are highly conserved, and which are degenerated. Profile hidden Markov models additionally store the likelihood of insertions and deletions to account for variation in protein length.

In essence, a pHMM is a linear chain of match, insertion, and deletion states (Figure 2.1). While match states represent the likelihood of seeing each amino acid at conserved positions, deletion states store the likelihood to skip a position. Insertion states allow for insertion of amino acids with respect to a match state. Thereby, pHMMs enable to calculate a probability for aligning a given amino acid to any position of the underlying multiple alignment. For a given sequence s, pHMMs employ statistical methods to calculate a probability that s was generated by the model. This probability is used to decide whether a sequence belongs to the represented family or not.

Pfam

The protein families database (Pfam) is a comprehensive collection of pHMMs for manually curated protein families (Finn *et al.*, 2006). The majority of these families represent *domains*, protein components with their own folding and molecular function, which mainly determine the function of proteins. The current version (version 21.0) consists of 8.957 families, matching about 74% of all known proteins (Finn *et al.*, 2006). Each family is represented by: a) a description of its functional role, b) a seed multiple alignment of representative family members, c) a global (ls) and a local (fs) pHMM, both constructed from the seed alignment, and d) a full multiple alignment comprising all members found in the public sequence databases using the pHMMs. Contrasting to the global model, the local pHMM does not penalize initial and terminal gaps, thus allowing partial matches to the pHMM to be found. In other words, with the fs pHMM a query sequence is *locally* aligned to the represented multiple alignment.

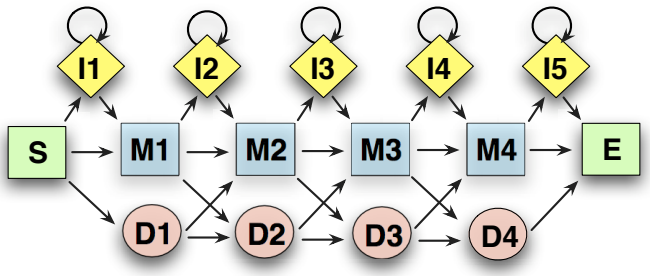


Figure 2.1: Simplified model architecture of a profile hidden Markov model (pHMM). The model parameters are estimated from a multiple alignment of representative members of a protein family. Each column of the alignment is represented either by a match and deletion or by an insertion state. Match states ($M_1, \ldots, M_4$) are assigned to conserved columns, i.e. columns where more than half of the sequences feature an amino acid. Match states represent the likelihoods for each amino acid to occur at a certain position. The likelihood of insertions and deletions is represented by insertion states $I_1, \ldots, I_5$ and deletion states $D_1, \ldots, D_4$. Moreover, a pHMM has a start S and termination state E representing the begin and end of an alignment. A given query sequence is aligned to a pHMM by assigning the most probable state transition to each of its amino acids (Durbin *et al.*, 1998).

2.3.2 Finding genes in prokaryotic genomes

Since the mid-1990s, gene finding software for bacterial genomes have become available allowing automated discovery of genes from raw genomic sequences (Borodovsky and MCIninch, 1993; Salzberg *et al.*, 1998; Frishman *et al.*, 1998; Badger and Olsen, 1999; Delcher *et al.*, 1999; Besemer and Borodovsky, 1999; Shmatkov *et al.*, 1999; Besemer *et al.*, 2001; Shibuya and Rigoutsos, 2002; Guo *et al.*, 2003; Larsen and Krogh, 2003; Delcher *et al.*, 2007). Subsequent developments have mostly focused on the introduction of novel techniques to capture sequence composition more accurately (Delcher *et al.*, 1999), modeling of the gene structure (Besemer *et al.*, 2001; Larsen and Krogh, 2003), and development of models that allow the discovery of multiple gene classes (Lukashin and Borodovsky, 1998; Mahony *et al.*, 2004).

In prokaryotes (bacteria and archaea), protein encoding genes are open reading frames (ORFs), which are sequences of codons beginning with a start codon, ending with a stop codon, and without internal stop codon. The task of gene prediction can therefore be regarded as a two class classification problem: Discriminating coding sequences (CDSs) from the majority of non-coding ORFs (nORFs), which correspond to genomic regions that are not transcribed and translated *in vivo*.

Two different approaches are applied for predicting protein encoding genes in prokaryotic genomes: *Intrinsic* and *similarity based* methods. Intrinsic methods, such as GLIMMER (Delcher *et al.*, 1999) or Genemark (Besemer and Borodovsky, 1999), analyze sequence properties of genomes to discriminate between coding sequences and non-coding ORFs. These methods exploit the different compositional properties

of coding and non-coding sequences, mainly caused by a bias in codon usage in the CDSs, which optimizes the translation efficiency in protein biosynthesis (Gouy and Gautier, 1982). Before intrinsic methods can be employed, they need to be trained to learn the specific sequence composition of the genome under study. Generative models with Markov properties (Durbin *et al.*, 1998), such as fixed-order Markov chains on nucleotides (Besemer *et al.*, 2001; Larsen and Krogh, 2003) or codons (Badger and Olsen, 1999), are often applied to represent the sequence composition. Additional sequence features, such as ribosome binding sites (Lewin, 2004) or overlaps between adjacent genes, can also be integrated into a probabilistic framework, if hidden Markov models are used to describe the context of a gene (Besemer *et al.*, 2001; Larsen and Krogh, 2003). Although these models provide a very detailed picture, their complexity can become problematic because of the high number of parameters that need to be adjusted during training (Larsen and Krogh, 2003).

Similarity based methods predict genes by searching for stretches of DNA that have been conserved during evolution. Several of these approaches discriminate conserved coding sequences from conserved non-coding regions based on their *synonymous substitution rate* (Badger and Olsen, 1999; Nekrutenko *et al.*, 2003a,b; Moore and Lake, 2003). These methods exploit that the genetic code is degenerated, i.e. most amino acids are encoded by several so called *synonymous codons*. Consequently, certain nucleotide substitutions can occur within a CDS without modifying the encoded amino acid sequence. These silent substitutions are called *synonymous*, otherwise *non-synonymous*. To maintain the amino acid sequence of the encoded protein, coding sequences show a much higher number of synonymous substitutions than non-coding sequences, which can be used to discriminate between them.

Because of the high accuracy initially reported for most gene finding programs, some might consider prokaryotic gene prediction as a solved matter, but from the point of a practitioner, this is not quite the case yet (e.g. Poole *et al.*, 2005). The development of techniques that improve predictions by combining the output of multiple programs (Tech and Merkel, 2003; McHardy *et al.*, 2004a) shows that accuracy can be increased. Moreover, the correct identification of short genes, genes with atypical sequence composition, and the prediction of genes where only little sequence material is available are still challenging problems for prokaryotic gene finders.

Short genes are generally more difficult to identify than longer ones because their sequences carry less information that can be evaluated for discrimination (Skovgaard *et al.*, 2001; Larsen and Krogh, 2003; Linke *et al.*, 2006). Additionally, the composition of genes is influenced by various factors, including the gene expression rate (McHardy *et al.*, 2004b), leading/lagging strand-related biases (Lafay *et al.*, 1999), or the transfer of genes between different bacterial species, called *horizontal gene transfer* (Smith *et al.*, 1992). Altogether, these influences lead to classes of genes with different sequence composition (Médigue *et al.*, 1991; Moszer *et al.*, 1999), which can be difficult to identify based on the evaluation of intrinsic sequence properties. For example, generative

methods, such as Markov chains or hidden Markov models, create a mean-based model of sequence composition by averaging over the sequence properties of their training collections. These methods have difficulties with genes that are best described by more than one distribution. This issue has been addressed by the inclusion of an additional model for genes with 'atypical' sequence composition (Lukashin and Borodovsky, 1998). On the other hand, short DNA sequences, such as plasmids, yield only small training sets for intrinsic sequence models. For complex models with many parameters this situation can lead to overfitting and a reduced prediction accuracy.

Gene finding in environmental metagenomes

Predicting protein encoding genes in metagenomes is problematic for several reasons. One problem is that assembled contigs may be too short to reveal the genome specific sequence properties, which is crucial in the application of intrinsic gene prediction methods. A second problem is the low sequence quality of assembled contigs, leading to sequence errors within the harbored CDSs. These errors may introduce insertions and deletions, entailing a partial shift of the codon sequence of a CDS (called *frame-shift*) as well as internal stop codons. In both cases, the affected CDSs are not ORFs anymore and hence may be missed by conventional, ORF-based gene finding methods. Although efforts have been made to overcome these problems – for example, the recently published METAGENE tackles the short contig length by estimating a GC content dependent usage of di-codons (Noguchi *et al.*, 2006) – most intrinsic gene finders are ineffective for environmental genomic sequences. As a consequence, the majority of CDSs in metagenomes are identified based on a BLAST search against databases of known proteins.

2.3.3 Reconstructing phylogenies

Phylogenetics is a field of biology concerned with the study of the evolutionary relationships of proteins or entire organisms, named *phylogenies*. These relationships are represented by phylogenetic trees, which in molecular biology are primarily derived upon the sequence characteristics of homologous proteins or genes (Zuckerkandl and Pauling, 1965). The basic terminology used to describe phylogenetic trees is introduced in Figure 2.2. Methods for inferring phylogenetic trees from molecular sequence data can be divided into two categories, *distance-based methods* and *discrete data methods*, also called *tree searching methods* (Baldauf, 2003). Discrete methods, e.g. parsimony or maximum likelihood (Felsenstein, 2004), examine each column of a multiple sequence alignment separately and search for the tree that best explains the observed sequences. While these methods provide a hypothesis for the evolution of every column in an alignment, their runtime complexity allows to analyze solely small data sets. Distance based methods, such as neighbor-joining (Saitou and Nei, 1987), are very fast and hence can be applied to data sets with hundreds of sequences, but require a precomputed distance matrix $D = (D_{ij})$. The distances D_{ij} between any two sequences s_i and s_j can for example be the proportion of identical characters in an alignment of s_i and s_j.

In the following, the *neighbor-joining* algorithm will be explained more thoroughly. For a tree T reconstructed from a distance matrix D, let n_i be the leaf in T that corresponds to sequence s_i. Then, T is correct with respect to D if for any two leaves n_i and n_j, the sum of branch lengths on the path from n_i to n_j equals D_{ij}. Furthermore, two nodes n_i and n_j are called *neighboring* nodes if the path between them has only one node.

If a matrix D is an *exact reflection* of a tree, i.e. a tree T exists that is correct in respect to D, then T is guaranteed to be reconstructed by the neighbor-joining algorithm (Studier and Keppler, 1988). Neighbor-joining neither presumes that sequences evolve at the same rate, nor makes any assumption about the rooting of a tree.

In essence, neighbor-joining iteratively joins nodes into larger and larger trees. Nodes that still need to be joined are stored in a list L of *active* nodes. Nodes that have been joined are marked as *inactive*. Initially, for each sequence s_i an active node n_i is added to L and the distance d'_{ij} between any two active nodes n_i and n_j is set to $d'_{ij} = D_{ij}$. In each successive step, the two active nodes that will be joined are determined according to the internal distance measure

$$d^*_{ij} = d'_{ij} - (r_i + r_j),$$

where r_i is the average distance of node n_i to all other active nodes

$$r_i = \frac{1}{N-2} \sum_{n_q \in L} d'_{iq}$$

and N is the total number of active nodes. One key concept of the neighbor-joining algorithm is that if D is the exact reflection of a tree T, then the two active nodes n_i and n_j with minimal distance d^*_{ij} are neighboring nodes in T. Accordingly, the algorithm selects the active nodes n_i and n_j with minimal d^*_{ij} and joins them to a new node n_k. The new node n_k is added to the list of active nodes L, the two selected nodes are removed from L. The distance d'_{kq} between n_k and each remaining active node n_q is set to

$$d'_{kq} = \frac{1}{2}(d'_{iq} + d'_{jq} - d'_{ij}).$$

This procedure is continued until only two active nodes are left, which are then joined by adding a branch between them.

2.3.4 Taxonomic classification of environmental DNA fragments

Inferring the source organisms (taxonomic origins) of environmental genomic fragments is one of the major questions in metagenomics and a challenging multiclass classification problem in bioinformatics. Since the pioneering work of Carl Woese and colleagues (Woese and Fox, 1977; Woese, 1987), 16sRNAs and 18sRNAs are commonly used to determine evolutionary relationships between organisms. Analogously, one type of strategy

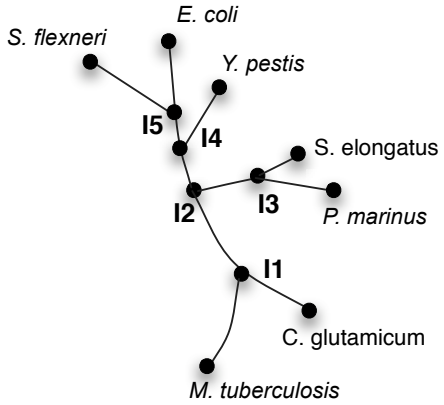


Figure 2.2: Sketch of a phylogenetic tree. The tree is composed of *nodes* and *branches* connecting nodes. Sequences from which the tree was reconstructed are represented by end nodes (*leaves*), labels indicate their source organisms, branch lengths reflect their evolutionary distance. The shown tree is *unrooted*, as it does not have an explicit root node representing the ultimate ancestor of all included sequences.

uses 16sRNA, 18sRNA, or other slowly evolving marker genes as 'phylogenetic anchors' to predict the taxonomic origins of environmental genomic fragments (Tringe and Rubin, 2005). While these methods frequently yield a high accuracy, only the fraction of fragments harboring such a marker gene can be taxonomically characterized. To overcome this limitation, novel methods have recently been devised that analyze the frequencies of short oligonucleotides or motifs to taxonomically classify genomic sequences (Teeling *et al.*, 2004; McHardy *et al.*, 2007). To the author's knowledge, however, none of these methods foster the classification of sequences shorter than 1000 bp and hence cannot be directly applied to metagenomes obtained by pyrosequencing. On the other hand, simply classifying fragments based on a best BLAST hit will only yield reliable results if close relatives are available for comparison (Koski and Golding, 2001).

2.4 Support Vector Machine learning technique

Support Vector Machines (SVMs) are a well studied and high performance supervised classification technique for two class (*binary*) classification problems (Boser *et al.*, 1992; Vapnik, 1995). In the past years, SVMs have increasingly drawn attention in bioinformatics (Noble, 2004), e.g. for the detection of protein family members (Jaakkola *et al.*, 2000; Leslie *et al.*, 2002; Hou *et al.*, 2003), RNA and DNA binding proteins (Cai and Lin, 2003), and for the functional classification of gene expression data for cancer detection (Furey *et al.*, 2000; Ramaswamy *et al.*, 2001). In the following, the key features of SVMs are introduced. First, the *maximum margin hyperplane* that optimally separates the items of two classes, second, the *soft margin hyperplane* that allows misclassification of outliers to increase the generalization ability, and third, *kernel functions* that enable to learn a separating hyperplane in a higher-dimensional feature space in order to achieve a non-linear classifier in the original input space.

Let v_j $(1 \leq j \leq m)$ be a set of training vectors with known class labels $y_j \in \{+1, -1\}$. Further, let $\mathcal{H}$ be a vector space with dot product $< x, x' >$, which in the context of SVMs is called *feature space*. Then, a *hyperplane* in $\mathcal{H}$ is given by a vector $w \in \mathcal{H}$ and a scalar $b \in \mathbb{R}^N$ and is defined as

$$\{x \in \mathcal{H} \mid < w, x > + b = 0\}$$

(Schölkopf and Smola, 2002). During training, SVMs explicitly learn a hyperplane (w, b), called *separating hyperplane*, which separates the vectors from the two training sets. The vector w that is learned by a SVM can be represented as a linear combination of weighted training vectors

$$w = \sum_{j=1}^{m} \alpha_j y_j v_j,$$

where α_j are weights that are assigned to each v_j during training (Schölkopf and Smola, 2002).

For a classification problem with linear separable classes, many separating hyperplanes exist. If we define the *margin* of a hyperplane as its minimal distance to any of the training vectors, then one key concept of SVMs is that the maximum margin hyperplane is learned. Selection of this 'optimal' separating hyperplane results in an improved accuracy for the classification of vectors with unknown class affiliations. With a learned hyperplane (w, b), a query vector v can be classified based on the *decision value*

$$d(v) = \sum_{j=1}^{m} \alpha_j y_j < v, v_j > + b \tag{2.1}$$

(Schölkopf and Smola, 2002). Given a learned hyperplane, a new vector v is commonly classified depending on the side of the hyperplane where it is located, i.e. whether $d(v)$ is larger than or smaller than 0.

To avoid overfitting, classifiers with a *soft margin* can be learned, which allow the misclassification of outliers during training. This is accomplished by bounding the weights α_j that are learned during training by a finite value C. As a consequence, the SVM parameter C influences the generalization ability of the learned classifier. If C is set to a small value, outlying training vectors are misclassified (Schölkopf and Smola, 2002) and herewith the margin of the hyperplane in respect to the remaining training vectors increased. This approach can for example be used to reduce overfitting in the case of small training sets. If C has a finite value, the resulting classifier is called a *soft margin SVM* (Schölkopf and Smola, 2002).

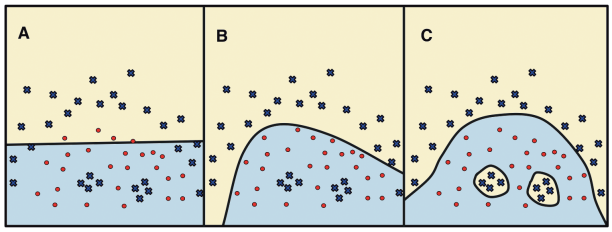


Figure 2.3: Class boundaries learned by a Support Vector Machine with different kernel functions. Circles and crosses represent instances of a toy example training set. Colored regions indicate the two classes learned by three example SVM applications. A) A linear decision function learned with a linear kernel $k(v, v_j) =< v, v_j >$. B) A polynomial kernel $k(v, v_j) = (< v, v_j > +r)^d$ allows realization of a polynomial separating surface. C) With a Gaussian kernel $k(v, v_j) = \exp\{-\gamma||v - v_j||^2\}$ the SVM can learn disjoint decision functions that surround a multitude of 'islands' of items from the same class (Hastie *et al.*, 2002).

For classification problems with classes that are not linearly separable in the original input space, a transformation into a suitable higher dimensional feature space may allow linear separation. This transformation can be performed implicitly if the dot product $< v, v_j >$ in Equation 2.1 is replaced by a kernel function $k(v, v_j)$. Kernel functions are distance measures that can be regarded as very time-effectively calculating the dot product in a higher dimensional feature space without explicit projection into that space (Schölkopf and Smola, 2002). By use of nonlinear kernel functions, complex and nonlinear decision functions can be learned by the SVM. The widely used *Gaussian kernel* $k(v, v_j) = \exp\{-\gamma||v - v_j||^2\}$ (Hastie *et al.*, 2002) allows a SVM to clearly separate vectors from two classes even if they are clustered in multiple disjoint subregions in the input space (Figure 2.3). The Gaussian kernel parameter γ influences the local behavior of the learned decision boundary. Setting a value for γ is a tradeoff between a well-fitted or more generalized decision boundary. A large value for γ results in irregular and noisy decision boundaries that are well fit to a training data set with more disjoint clusters. Small values for γ, on the other hand, result in smooth and stable boundaries that avoid overfitting and are more robust (Hastie *et al.*, 2002).

2.5 Measures of accuracy

In the presented thesis, classification techniques are employed for predicting protein encoding genes as well as for inferring the taxonomic origins of environmental genomic fragments. While the first task can be regarded as a binary classification problem (Section 2.3.2), the second has multiple classes (Section 2.3.4). For binary classification problems, the accuracy of a classifier can be measured by the sensitivity, specificity, false positive rate, and false negative rate (Baldi and Brunak, 2001; Baldi *et al.*, 2000). In this study, these measures are adapted to multiclass problems. Assume a classification problem with two or more classes – in the context of this work CDSs and nORFs or several taxonomic categories – and a test set of items with known class labels. Then, the accuracy of a classifier can be measured by comparing the predicted with the known class label of each item from the test set.

For simplicity, let the i-th class be denoted as *class i*. Further, let P_i be the total number of items from class i; TP_i the number of items that is correctly assigned to class i; FP_i the number of items that is erroneously assigned to class i; TN_i the number of items from any class $j \neq i$ that is not assigned to class i; and FN_i the number of items from class i that is misclassified into some class $j \neq i$. Further, let U_i be the number of items from class i that cannot be assigned to any class. Note that $P_i = TP_i + FN_i + U_i$. Then, the *sensitivity* measures the proportion of items that is correctly classified. For a class i, it is defined as $Sn_i = \frac{TP_i}{P_i}$. The *specificity* measures the reliability of classifications and is defined as $Sp_i = \frac{TP_i}{TP_i+FP_i}$. The *false negative rate* is defined as $FNrate_i = \frac{FN_i}{P_i}$. It measures the proportion of items from class i that is falsely assigned to any class $j \neq i$. The *unknown rate* measures the proportion of items that cannot be classified and is defined as $Urate_i = \frac{U_i}{P_i}$. The *false positive rate* is the proportion of items from any class $j \neq i$ that is falsely assigned to class i. It is defined as $FPrate = \frac{FP_i}{FP_i+TN_i}$.

In this work, the accuracy of different gene prediction programs was measured only for the CDS class. In the context of gene finding, the sensitivity, specificity, and false positive rate therefore always refer to these measures for the CDS class. Furthermore, as ORFs are always classified as either CDS or nORF, both U_i and $Urate$ equal 0. The *correlation coefficient* was used to characterize the *overall accuracy* of different gene finding programs. It is defined as

$$cor = \frac{N * Sn_{CDS} * Sp_{CDS} - TP_{CDS}}{\sqrt{(N * Sn_{CDS} - TP_{CDS}) * (N * Sp_{CDS} - TP_{CDS})}},$$

where $N = TP_{CDS} + FP_{CDS} + TN_{CDS} + FN_{CDS}$. It describes the agreement of assigned class labels and known class labels with a single value in the range of [-1,1].

2.6 Measuring the biodiversity of microbial communities

Several descriptors are used to characterize the biodiversity of complex communities. These descriptors not only condense and summarize the wealth of information about the number and relative abundance of species, but further provide quantitative measures that can be used to compare communities.

Two important community attributes are *richness* and *evenness*. While the richness is simply the total number of species present, the evenness measures the relative commonness and rarity of organisms (Morin, 1999). In metagenomics, measuring the species richness is problematic if the sample size was not sufficient enough to cover the complete range of species present. The overall diversity of an environmental sample, including both richness and evenness, can be characterized with *Shannon's diversity index* (Shannon and Weaver, 1963).

In the context of this work, for a taxonomic rank r out of phylum, class, order, and genus, Shannon's diversity index is defined as

$$H' = -\sum_{i=1}^{K} p_i \ln p_i,$$

where p_i is the fraction of genomic fragments from a metagenome to which the i-th taxon of rank r was assigned and K is the number of taxa at rank r. The *species evenness* can then be defined as

$$J = \frac{H'}{\ln(H_{max})},$$

where H_{max} is the total number of taxa found at rank r. Note, that the diversity and evenness are usually measured at the rank of species. Nonetheless, in this study, diversity and evenness are measured at the rank of phylum, class, order, and genus, as quantitative species information is frequently not available for metagenomic data sets.

CHAPTER 3

Contributions

This chapter offers a synopsis of the contributions made by the author in the analysis of whole genomes and metagenomes within the presented thesis. For this purpose, a selection of six scientific articles is included, four published papers (Paper I-IV) and two submitted manuscripts (Paper V and VI):

Paper I: L. Krause, A. C. McHardy, T. W. Nattkemper, A. Pühler, J. Stoye, and F. Meyer. GISMO – gene identification using a Support Vector Machine for ORF classification. *Nucleic Acids Res*, 35:540–549, 2007.

Paper II: A. Krause, A. Ramakumar, D. Bartels, F. Battistoni, T. Bekel, J. Boch, M. Böhm, F. Friedrich, T. Hurek, L. Krause, B. Linke, A. C. McHardy, A. Sarkar, S. Schneiker, A. A. Syed, R. Thauer, F.-J. Vorhölter, S. Weidner, A. Pühler, B. Reinhold-Hurek, O. Kaiser, and A. Goesmann. Complete genome of the mutualistic, N2-fixing grass endophyte *Azoarcus sp.* strain BH72. *Nat Biotechnol*, 24:1385–1391, 2006.

Paper III: R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. The subsystems approach to genome annotation and its use in the Project to Annotate 1000 Genomes. *Nucleic Acids Res.*, 33:5691–5702, 2005.

Paper IV: L. Krause, N. N. Diaz, D. Bartels, R. A. Edwards, A. Pühler, F. Rohwer, F. Meyer, and J. Stoye. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics*, 22:e281–e289, 2006.

Paper V: F. Rohwer, H. Liu, F. E. Angly, S. Rayhawk, L. Krause, R. Olson, B. Brito, R. Stevens, and R. A. Edwards. Releasing metagenomics data. *Submitted*.

Paper VI: L. Krause, N. N. Diaz, A. Goesmann, F. Rohwer, S. Kelley, R. A. Edwards, and J. Stoye. Taxonomic classification of short environmental DNA fragments. *Submitted*.

This chapter consist of two sections, the first (3.1) deals with contributions made in improving the knowledge of the gene content of prokaryotic genomes and the second (3.2) considers novel methods for analyzing and characterizing metagenomic data sets. In each section, a summary of the corresponding publications is provided together with relevant complementary information about the developed algorithms as well as information about their applications in whole-genome and metagenome projects. Please note that in the scope of this chapter 'gene' always refers to protein encoding genes.

3.1 Improving annotation of prokaryotic genomes

Genome annotation is an issue of crucial importance, approximately 1000 microbial genome projects are currently in progress (according to http://www.genomesonline.org/, as of May, 2007). In addition, various genome re-annotation efforts are undertaken to improve knowledge of the genes present therein and their cellular functions. For this gigantic amount of genomic data it is essential to rely on accurate methods for the automated identification and functional characterization of genes, allowing a reduction of the manual annotation effort while at the same time achieving a high quality.

In the following, contributions made by the author in improving the gene content knowledge of prokaryotic genomes are described. After the introduction of GISMO, a highly accurate gene finder for prokaryotic genomes, the genome project of *Azoarcus sp.* is presented, in which the author identified missing genes and participated in the functional genome annotation. Finally, the subsystems approach is delineated, a high-throughput annotation strategy.

3.1.1 GISMO – gene identification using a Support Vector Machine for ORF classification.

L. Krause, A. C. McHardy, T. W. Nattkemper, A. Pühler, J. Stoye, and F. Meyer. *Nucleic Acids Res*, 35:540–549, 2007.

The open-source, prokaryotic gene finder GISMO, introduced in Paper I, combines searches for protein domain families with composition-based classification of ORFs using a Support Vector Machine. Initially, a similarity based search for conserved domain families is conducted using Pfam profile hidden Markov models (pHMMs). Subsequently, a Support Vector Machine is trained to learn the genome specific sequence composition of coding sequences (CDSs) and non-coding ORFs (nORFs). Each ORF is represented as a 64-dimensional vector of codon frequencies. For training, ORFs with significant Pfam hits are used as training set for the CDS class, ORFs located on a different frame within the 'shadow' of these (overlapping by 90bp or more) as the training set for the nORF class. During training, a hyperplane is learned in a higher dimensional feature space that optimally separates ORFs from the two classes. Finally, the Support Vector Machine is used to classify all ORFs of a genome into coding and non-coding based on their sequence composition.

Using GISMO, almost the complete gene content of bacterial and archaeal genomes can be elucidated with high reliability. In a performance evaluation on 165 genomic sequences, the algorithm identified on average 94.3% of the known genes and 94.3% of the predictions corresponded to a known gene. GISMO even detected 98.9% from the subset of the more reliable genes associated with a functional description in the annotations. A high accuracy was also achieved for short genes (Paper I, Table 3), horizontally transferred genes (Paper I, Table 4), and short DNA sequences such as plasmids (Paper I, Table 5). A large scale comparative analysis of the additionally GISMO predictions for 165 genomes showed that probably a significantly high number of genes are currently missing in the public genome annotations (e.g. Paper I, Figure 2 and Figure 5).

Related work

Within the presented thesis, GISMO has been applied in more than 20 international genome projects, for both *de-novo* annotation of genomes as well as in re-annotation projects to refine the gene content of organisms. Furthermore, GISMO has been extensively used in the international effort to 'Annotate a Thousand Genomes', as well as in a collaboration with Jürgen Hausmann, Bavarian Nordic GmbH, for identifying the gene content of the chorioallantois vaccinia virus Ankara. The results for the latter study are part of the submitted manuscript:

> C. Meisinger-Henschel, M. Schmidt, S. Lukassen, B. Linke, L. Krause, S. Konietzny, A. Goesmann, P. Howley, P. Chaplin, M. Suter, and J. Hausmann. The genomic sequence of chorioallantois vaccinia virus Ankara, the ancestor of modified vaccinia virus Ankara. *Submitted*.

3.1.2 Complete genome of the mutualistic, N2-fixing grass endophyte *Azoarcus sp.* strain BH72

A. Krause, A. Ramakumar, D. Bartels, F. Battistoni, T. Bekel, J. Boch, M. Böhm, F. Friedrich, T. Hurek, L. Krause, B. Linke, A. C. McHardy, A. Sarkar, S. Schneiker, A. A. Syed, R. Thauer, F.-J. Vorhölter, S. Weidner, A. Pühler, B. Reinhold-Hurek, O. Kaiser, and A. Goesmann. *Nat Biotechnol*, 24:1385–1391, 2006.

In the genome annotation of the kallar grass symbiont *Azoarcus sp.* strain BH72, the author contributed by means of both, using GISMO to search for genes that were missing in the annotation, as well as in the functional characterization of its gene content. *Azoarcus sp.* is an important model organism for studying mechanisms of host-symbiont interactions, as it supplies its host with nitrogen without eliciting disease symptoms. Comparative analyses with close relatives not associated with plants revealed genes that may be beneficial for the symbiotic lifestyle. In particular, the *Azoarcus sp.* comparative study elucidated genes that may play an important role for interacting with the host as well as a lack of genes associated with plant pathogens, such as genes encoding toxins or enzymes degrading plant cell walls. These findings may aid in developing agrobiotechnological applications of nitrogen supplying bacteria.

3.1.3 The subsystems approach to genome annotation and its use in the Project to Annotate 1000 Genomes.

R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. *Nucleic Acids Res.*, 33:5691–5702, 2005.

In cooperation with Ross Overbeek, Fellowship for the Interpretation of Genomes, GISMO has been extensively applied in the international project to 'Annotate a Thousand Genomes' – aiming at creating high-quality annotations for all thousand genomes that will be available within the coming months. For the functional annotation of genes, the central principle of this effort is to define so called *subsystems*, sets of genes that together implement a specific biological process or structural complex, and the consecutive annotation of single subsystem over the entire collection of publicly available genomes by experts.

Compared to the traditional strategy where a genome is annotated 'gene-by-gene', the subsystems approach has several key advantages: a) genes that make up a subsystem are annotated by an expert on that subsystem, b) it is more effective to annotate gene families in the context of related genes families than to annotate each gene in isolation (Osterman and Overbeek, 2003), and c) genes missing in the current annotations are easily identified. The collection of annotated subsystems is made available to the research community via a clearinghouse. Altogether, subsystems provide manually curated annotations of high quality and a central resource for studying the gene content and metabolic capacities of organisms. An essential prerequisite for obtaining high-quality annotations is an accurate gene prediction, which in the context of this project has been accomplished using GISMO in combination with GLIMMER, CRITICA, SearchforRNAs (N. Larsen, unpublished), and tRNAscan-SE (Lowe and Eddy, 1997).

3.2 Novel methods for characterizing microbial metagenomes.

Metagenomics lends striking insights into the species composition and metabolic activities of natural microbial communities. But, while the gene content of whole genomes can nearly completely be revealed in an automated manner and the majority of identified genes are routinely functionally characterized based on sequence similarity, the computational analysis of metagenomes is still in its infancy. The analysis of mixtures of short DNA fragments of unknown origin and with divergent sequence properties, combined with a low sequence quality is loaded with difficulties, thus making the application of many existing tools ineffective.

This section summarizes contributions in characterizing environmental metagenomes. In Paper IV a novel algorithm is presented for predicting genes in short environmental contigs of low sequence quality. Furthermore, in a cooperation with Robert Edwards and Forest Rohwer, San Diego State University (SDSU), CA, methods were developed for directly identifying gene fragments in the un-assembled reads of a metagenome, followed by their categorization into functional and taxonomic groups. The developed strategy as well as precomputed results for several samples were made available to the scientific community as part of Paper V. The taxonomic classification of short gene fragments was further improved, as described in Paper VI.

3.2.1 Finding novel genes in bacterial communities isolated from the environment.

L. Krause, N. N. Diaz, D. Bartels, R. A. Edwards, A. Pühler, F. Rohwer, F. Meyer, and J. Stoye. *Bioinformatics*, 22:e281–e289, 2006.

Environmental contigs are a valuable resource for finding novel (habitat-specific) genes that are missed by cultivation-dependent methods. Nevertheless, currently in the majority of metagenome projects BLAST searches against a databases of known proteins are conducted to identify genes. This strategy completely fails to detect genes in the absence of a sufficient similarity to any of the sequences stored in the database used for comparison.

On the other hand, most of the existing intrinsic methods are ineffective when directly applied to metagenomes. For example, GISMO can inherently predict genes in short genomic sequences and accurately discriminate between non-coding and coding sequences with divergent compositional properties, as stated above. Nonetheless, also the SVM-based prediction of genes has been impractical for contigs shorter than 10.000 bp, because they do not provide a sufficient training set. Additionally, the prediction accuracy is strongly affected by a low sequence quality, because GISMO, as almost all existing prokaryotic gene finders, fails to identify genes with frame-shifts or in-frame stop codons.

In Paper IV, a novel gene finding algorithm is presented which incorporates features that help to overcome the following problems: a) the short length, b) contained in-frame stop codons, and c) frame-shifts of assembled contigs from environmental samples. For the identification of genes with none or only weak sequence similarity to those present in the public sequence databases, the central principle of the algorithm is to search for conserved regions within a metagenome.

Initially, the algorithm searches for conserved regions based on a BLAST search of each contig of a metagenome against the entire metagenome itself. Next, in order to discriminate between conserved coding sequences and conserved non-coding regions and to additionally identify gene boundaries, different sequence features are taken into account: a) the synonymous substitution rate for each aligned nucleotide of a contig, b) the positions of stop codons in the contig, c) the position of stop codons in matching sequences, and d) the end of BLAST hits. These features are used to score the potential of each aligned nucleotide of a contig to be coding in each of the six reading frames (Paper IV, Figure 1). Based on these scores, a dynamic programming algorithm is employed to find the best partitioning of a contig into coding and non-coding regions (Paper IV, Figure 2).

A performance evaluation on a synthetic metagenome showed that by searching for sequence similarities within a metagenome, the algorithm is capable of detecting a high fraction of its gene content, depending on the species composition and the overall size of the original sample (Paper IV, Table 2). On the whole, the implemented strategy is very robust against the most common problems encountered when predicting genes in environmental contigs and more importantly, it will identify genes regardless of whether these have a known homolog or not.

3.2.2 Releasing metagenomics data

F. Rohwer, H. Liu, F. E. Angly, S. Rayhawk, L. Krause, R. Olson, B. Brito, R. Stevens, and R. A. Edwards. *Submitted.*

Nowadays, sequencing of environmental samples is becoming a routine process, allowing individual laboratories to produce giga-basepair data sets. The main bottleneck in studying metagenomes is their computational analysis because of several reasons, including a lack of computing resources, software that can handle metagenomic data sets, or simply expertise. Paper V describes the SDSU Center for Universal Microbial Sequencing (SCUMS) Web site, which delineates strategies for analyzing metagenomes and contains different software tools for their characterization. Moreover, it makes more than 100 environmental sequence data sets available to the research community, complemented by precomputed results. The available data sets were mainly obtained using pyrosequencing, accompanied by several metagenomes obtained by the Sanger technique. This is the largest released collection of environmental sequences isolated from different environments, allowing researchers to readily and easily study metagenomes from disparate habitats.

CARMA – a pipeline for characterizing short-read metagenomes

When applied to environmental samples, the 454 pyrosequencing technique enables sequencing of large libraries at a low cost, but the short length of the reads obtained poses a crucial challenge for their consecutive computational analysis. For example, the assembly of reads into contigs becomes extremely difficult and standard analysis steps – such as BLAST searches against databases of known proteins – yield only unreliable results. As part of the presented thesis, the pipeline CARMA was developed, for analyzing short-read metagenomes obtained by 454 pyrosequencing, without a prior assembly step (Figure 3.1). Following an initial gene finding phase, the genetic diversity and taxonomic composition of the underlying microbial communities are characterized. The developed strategy as well as precomputed results for several environmental samples are part of the SCUMS Web site.

The development and evaluation of GISMO showed that profile hidden Markov models are very accurate means for detecting short coding sequences. Thus, in an initial step, CARMA employs pHMMs from the Pfam database to identify gene fragments, called environmental gene tags (EGTs, see Section 2.2 on page 6), in the unassembled reads of a metagenome. In detail, all reads of a sample are translated into each of the six reading frames. Next, reads are locally aligned to each Pfam family using its local pHMMs. Regions with a hit of E-value < 0.01 are predicted as EGT. Herewith, the described method complements GISMO and the gene finding algorithm developed for environmental contigs. While GISMO accurately identifies genes in prokaryotic genomes, the latter enables gene identification in environmental contigs. On the other hand, CARMA provides the opportunity to search for gene fragments in environmental sequences with length as short as 100 bp.

For characterizing the genetic diversity of the underlying community, EGTs are categorized into functional genetic groups according to the Gene Ontology (GO), which provides a controlled vocabulary to describe the molecular function of proteins (Ashburner *et al.*, 2000). For each EGT, GO-terms are obtained from its matching Pfam family description. By comparing the GO-term profiles of different samples, gene functions that are important for survival in a habitat or adaptation to a niche are identified. On the one hand, GO-term profiles are visualized in a heat map (Figure 3.2), enabling researches to rapidly browse and compare the potential metabolic and cellular activities of different communities. Moreover, for a given set of metagenomes, significantly overrepresented GO-terms in each of them are identified in a pair-wise comparison of the GO-term profiles of all metagenomes versus all metagenomes. To decide if a GO-term is significantly overrepresented, a *G*-test is employed with the null hypothesis $H0$: There is no difference in the abundance of identified EGTs between the two samples to which the GO term was assigned. For details on the *G*-test refer to Fowler *et al.* (1998).

To characterize the taxonomic composition of the underlying communities, the potential source organism of each EGT g is inferred as follows: First, it is compared (aligned) to each member of its matching Pfam family. Next, the longest common prefix of the taxa of all members having a sequence identity of at least $95\% * I_{max}(g)$ is assigned to g, where the *sequence identity* is the fraction of identical amino acids in an aligned region and $I_{max}(g)$ is the maximal sequence identity obtained in the aforementioned comparison. As an example, 'Bacteria Cyanobacteria' is the longest common prefix of the two taxa

- 'Bacteria Cyanobacteria Prochlorales Prochlorococcus'
- 'Bacteria Cyanobacteria Chroococcales Synechococcus'

With this approach, for each metagenome a taxonomic profile is obtained, reflecting the taxonomic composition of the underlying microbial community (Figure 3.1). Trends in the composition of different communities can be unveiled in a comparative analysis of their profiles.

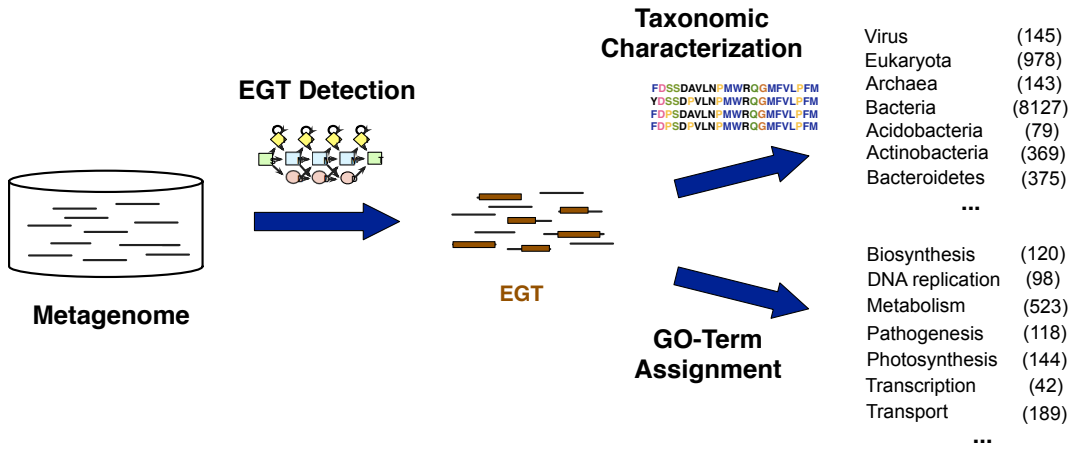


Figure 3.1: Overview of CARMA, a pipeline for characterizing short-read metagenomes. Initially, gene fragments (environmental gene tags, EGTs) are identified using Pfam profile hidden Markov models. Subsequently, a taxonomic origin and a GO-term is assigned to each identified EGT, yielding a community-specific genetic and taxonomic profile. The number of EGTs to which each taxonomic category and GO-term was assigned is depicted in parenthesis.

Related work

The potential of CARMA to provide deep insights into the genetic diversity and taxonomic composition of environmental samples has already been demonstrated in several practical applications. For example, it has been applied to unveil habitat-specific metabolic and taxonomic traits in four coral reef microbial communities, in a collaboration with Elisabeth Dinsdale, Robert Edwards, and Forest Rohwer (San Diego State University, CA). The results indicated that human disturbance has a dramatic impact on the coral reef microbiome (Dinsdale *et al.*, *submitted*). Furthermore, CARMA has been used for studying the diversity of coral reef viral assemblages, in a cooperation with Stuart Sandin (Center for Marine Biodiversity and Conservation Scripps Institution of Oceanography, San Diego, CA), and for characterizing a plasmid sample isolated from a wastewater treatment plant, in collaboration with Andreas Schlüter (Bielefeld University, Germany). In the latter study, the genetic characterization revealed a high proportion of protein families involved in plasmid replication, mobilization, and plasmid stability. Furthermore, fragments of genes promoting virulence and resistance were elucidated. Altogether, the derived results indicated that plasmids residing in the bacterial community of wastewater treatment plants harbor a high genetic diversity.

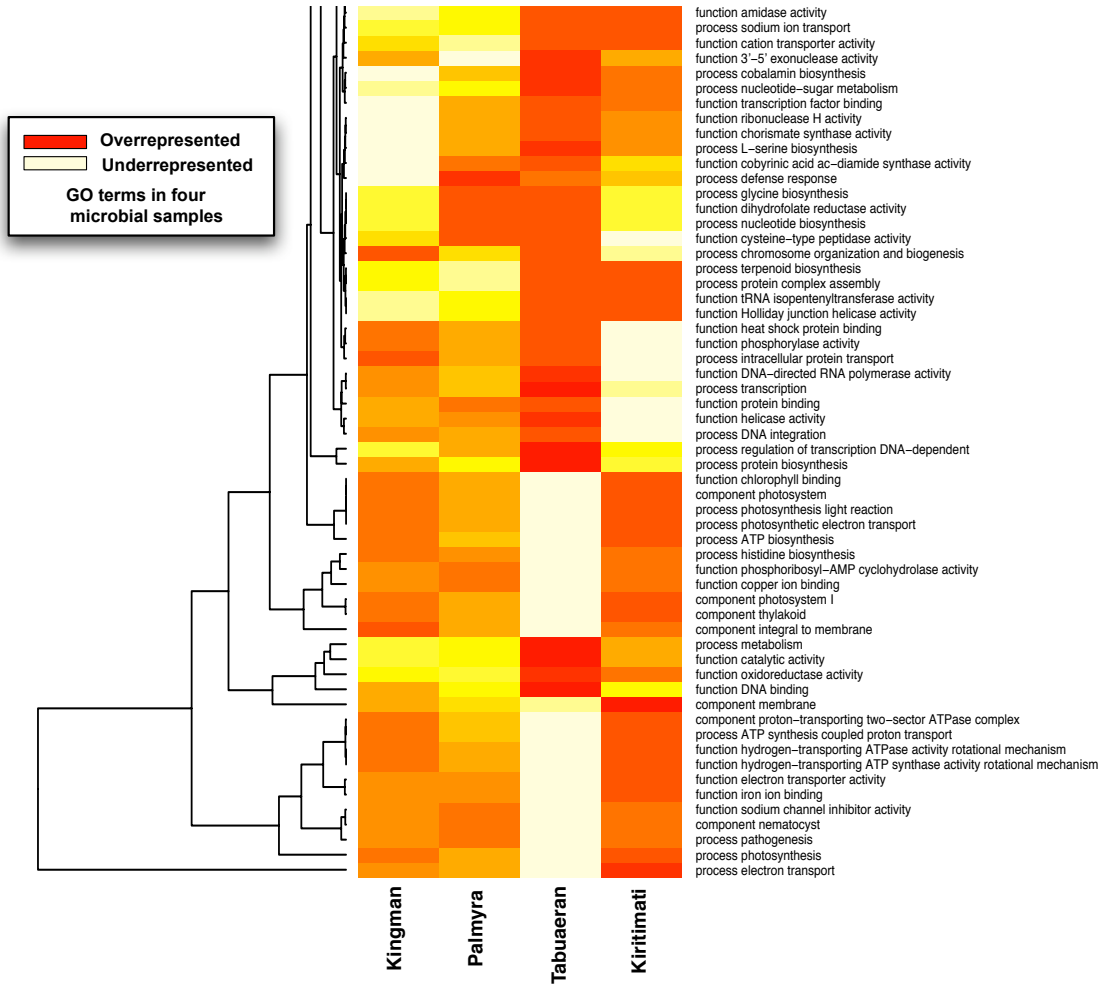


Figure 3.2: GO-term profiles of four microbial metagenomes isolated from Kingman, Palmyra, Tabuaeran, and Kiritimati coral reefs. The colors illustrate the abundance of each GO-term in each metagenome.

The results for the coral reefs microbial samples and for the coral reefs viral samples are part of two submitted manuscripts:

1. E. A. Dinsdale, O. Pantos, S. Smriga, R. A. Edwards, F. E. Angly, D. Hall, E. Brown, M. Haynes, L. Krause, E. Sala, S. A. Sandin, R. V. Thurber, B. L. Willis, F. Azam, N. Knowlton, and F. Rohwer. Diversity and disturbance on coral reefs. *Submitted*.

2. S. A. Sandin, N. Knowlton, F. E. Angly, E. E. DeMartini, R. A. Edwards, A. M. Friedlander, J. B. C. Jackson, L. Krause, J. E. Maragos, D. Obura, F. Rohwer, E. Sala, P Salamon, and J. E. Smith. Dramatic changes in microbial communities on coral reefs. *Submitted*.

3.2.3 Taxonomic classification of short environmental DNA fragments

L. Krause, N. N. Diaz, A. Goesmann, F. Rohwer, S. Kelley, R. A. Edwards, and J. Stoye. *Submitted.*

The taxonomic composition of short-read metagenomes can be characterized by inferring the taxonomic origins of gene fragments (environmental gene tags, EGTs), as illustrated in the previous subsection. While CARMA roughly estimates the source organisms of EGTs using a simple decision rule, Paper VI presents a novel algorithm for the taxonomic classification of EGTs based on a phylogenetic analysis.

In the first phase, EGTs are identified in the unassembled reads of a metagenome using Pfam profile hidden Markov models (pHMMs), analogous to the gene prediction strategy used in CARMA. In the second phase, a phylogenetic tree is reconstructed for each matching Pfam family using the neighbor-joining method. Taxonomic origins are assigned to EGTs based on their location in the respective family tree.

The algorithm accurately classifies EGTs as short as 33 amino acids ($\approx$ 100 bp) up to the rank of genus (Paper VI, Figure 2 and Figure 3). For EGTs from taxa that are well represented in Pfam (archaea, bacteria, eukaryota, and viruses, 20 phyla, 27 classes, 59 orders, 69 genera) between 97% (superkingdom) and 68% (genus) of the predicted taxa are correct. The sensitivity on the other hand ranges from 90% for superkingdom to 40% for the rank of genus. Also for poorly represented taxa, reliable predictions are obtained, with a specificity ranging from 84% for superkingdom to 65% for genus.

Within Paper VI, the algorithm was applied in a comparative study of three metagenomes obtained by 454 pyrosequencing. Despite the short length of the genomic sequence fragments, the analysis clearly revealed substantial differences in the taxonomic composition, species diversity, and species evenness of microbial communities sampled from different aquatic environments (Paper VI, Table 2 and Figure 7).

CHAPTER 4

Conclusions

Fueled by advances in sequencing techniques, sequence analysis is still an important field of research in computational biology. In this thesis, strategies were devised for predicting genes in whole genomes and environmental genomic contigs. These approaches are complemented by methods for identifying gene fragments in metagenomes obtained by ultra-fast 454 pyrosequencing, despite that this technique produces only short sequence reads with an averaging length of 100 bp. Identified gene fragments are subsequently categorized into taxonomic and functional groups in order to appraise the species composition and metabolic capacities of the underlying communities. Software developed in the presented thesis has been applied in various genome and metagenome projects, contributed in studying microbes with relevance to human health, agriculture, environmental pollution, coral health, biotechnological applications, and many other areas of microbiology.

Support Vector Machines (SVMs) in combination with Pfam profile hidden Markov models give the capacity to unveil almost the complete gene content of prokaryotic genomes in a fully automated manner. In particular, SVMs are well suited and easy to handle for the task of gene identification in prokaryotic genomes. They can learn to optimally discriminate protein encoding genes (CDSs) from non-coding ORFs (nORFs) – combined with a Gaussian kernel even if CDSs and nORFs are distributed over multiple disjoint clusters in the input space. This property is convenient for gene prediction. The sequence composition of coding sequences is influenced by various factors affecting different genomes to different extents, and for each case the optimally separating boundaries can be found anew. As SVMs further avoid over-fitting by learning a soft margin hyperplane, they are able to accurately predict genes for short genomic sequence, which yield only small training sets to learn their specific sequence composition. Profile hidden

Markov models on the other hand are highly accurate in detecting conserved coding sequences based on sequence similarities to Pfam protein families. Since the majority of Pfam families models protein domains and not whole proteins, this approach enables detection of genes with protein domains occurring in different order than in known proteins.

A high fraction of genes in metagenomic contigs can be elucidated in an automated manner, based on a search for regions that are conserved within a metagenome. In this thesis, a gene finding algorithm is presented that integrates information on conserved regions and different sequence properties and finally identifies the optimal partition of a contig into genes and intergenic regions using dynamic programming. With this strategy, the main challenges encountered when predicting genes in metagenomes can be handled, including short contig length and low sequence quality. Thus, metagenomic contigs are a valuable resource for hunting novel genes missed by culture dependent methods as well as for in-depth studies of the gene content of natural microbial communities.

The ultra-fast pyrosequencing technique in combination with methods developed in this thesis gives the capacity to rapidly and cost-effectively assay natural microbial communities, without a bias that cloning or assembly procedures may possibly introduce. Despite that only short gene fragments are identified, these snippets can accurately be functionally and taxonomically characterized, yielding a community-specific genetic and taxonomic profile. Profiles of unassembled metagenomes can in turn be used for quantitative studies of the gene content and species composition of the underlying communities. Based on a comparative analysis, important traits in the metabolic activities, cellular processes, and taxonomic composition of each community are unveiled. Furthermore, gene functions can be elucidated that are important for survival in the environment where a metagenomic sample is taken or for adaption to a niche. Altogether, profiling approaches enable to answer two substantial questions in microbial ecology: First, which microbes live in a certain environment, and second, what are they doing by means of their metabolism and cellular processes.

CHAPTER 5

Future directions

As of to date, almost the complete gene content of prokaryotic genomes can be automatically revealed and efforts like the subsystems approach promise to provide high-quality functional annotations. These methods will pave the way for high-throughput comparative analysis of microbial genomes. While during the last decade genome projects mainly aimed at annotating single genomes, driven by recent advances in sequencing technologies, at present many projects are initiated to sequence and study various strains of the same or closely related species. These approaches enable to identify genotype-phenotype associations, for example genes and mutations promoting pathogenesis or encoding toxins. However, new demands arise for computational biology, such as comparative genomics methods to trace a certain phenotype back to variations in the genomic sequence. One impressive application of these approaches was a study of 169 avian influenza isolates, carried out by Obenauer and colleagues (Obenauer *et al.*, 2006). Genomic variations likely to promote virulence could be revealed, that are promising targets for novel antiviral therapies.

In the emerging field of metagenomics, preliminary important steps have been accomplished in studying the ecology of free-living microbes, but still much work remains to be done. Estimates on 16sRNAs suggest that only 1% of microbes can be cultivated with conventional methods (Hugenholtz, 2002). In this regard, we would ideally like to reconstruct complete genomes from metagenomic data sets. But at present, the main limiting factor for this task is simply the cost of obtaining a sufficient sequence coverage of organisms present in an environment. Additionally, the assembly of reads from a mixture of species with unbalanced abundances and inter-species genetic heterogeneity is a challenging venture. In the near future, both limitations may be circumvented by single molecule sequencing techniques, which promise to sequence long stretches of genomic DNA at once, without a prior fragmentation step (Dear, 2003; Chan, 2005).

Despite the impressive progress in microbiology during the past decades, metagenomic studies continuously reveal how little is known about the composition, diversity, interactions, and dynamics of natural microbial communities. So far, mainly static snapshots of the composition and potential metabolism of the microbiota of a few selected environments were obtained. In order to explore the entire microbial diversity, at present several large scale projects are in the process of being initiated to sequence the microbiomes of various disparate environments (Pennisi, 2007). Particularly the study of the human microbiota, which is estimated to contain several orders of magnitude more genes than the human genome, could reveal how microorganisms contribute to our health and disease. For example, using metagenomic approaches, Gill *et al.* (2006) could gain insights into the crucial role of the distal gut microbiome in human biology, which once more makes clear that humans are a 'superorganism whose metabolism represents an amalgamation of microbial and human attributes'.

In this thesis, it has been demonstrated that profiling methods can provide deep insights into the species composition and genetic diversity of microbial communities. To identify genes that are important to survival in a certain environmental condition and to further identify phenotype-genotype associations, databases are required to store attributes of the environment from which metagenomes are taken. These attributes should for example include location, temperature, and chemical properties. At present, attempts are being made by the Genomic Standards Consortium to define standardizations to represent this type of 'metadata' (Field *et al.*, 2005, 2006). Furthermore, biomarkers need to be determined that allow to rapidly detect certain groups of organisms or gene functions of interest, including pathogens or genes transmitting resistance. This could for example aid in applications such as food quality control, diagnosing the 'health state' of environments, or protecting against bioterrorism.

If our goal is to completely understand the interactions, dynamics, response to environmental changes, and biochemical and physiological processes of microbial communities, we need to launch more in depth studies of their metagenomes, and monitor changes in communities over time. Additionally, metagenomic studies should be complemented by exploring the community gene transcription (Gentry *et al.*, 2006; Gao *et al.*, 2007) and proteome (Ram *et al.*, 2005) and by scrutinizing epigenetic modifications associated with certain environmental conditions. In the long term, our goal is to model the biochemical processes and dynamics of entire microbial ecosystems, allowing to make predictions about the effects of perturbations of environmental conditions, including pollution, drug treatment, the release of transgenic organisms, or climate change.

Bibliography

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, (1997).

Angly F. E., Felts B., Breitbart M., Salamon P., Edwards R. A., Carlson C., Chan A. M., Haynes M., Kelley S., Liu H., Mahaffy J. M., Mueller J. E., Nulton J., Olson R., Parsons R., Rayhawk S., Suttle C. A., Rohwer F.: The marine viromes of four oceanic regions. *PLoS Biol*, 4:e368, (2006).

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., Sherlock G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25–29, (2000).

Badger H., Olsen G. J.: CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol*, 16:512–524, (1999).

Baldauf S. L.: Phylogeny for the faint of heart: a tutorial. *Trends Genet*, 19:345–351, (2003).

Baldi P., Brunak S.: *Bioinformatics The Machine Learning Approach*. The MIT Press, Cambridge, MA (2001).

Baldi P., Brunak S., Chauvin Y., Andersen C. A., Nielsen H.: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, (2000).

Besemer J., Borodovsky M.: Heuristic approach to deriving models for gene finding. *Nucleic Acids Res*, 27:3911–3920, (1999).

Besemer J., Lomsadze A., Borodovsky M.: GeneMark.S: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*, 29:2607–2618, (2001).

Bohannon J.: Metagenomics. Ocean study yields a tidal wave of microbial DNA. *Science*, 315:1486–1487, (2007).

Borodovsky M. Y., MCIninch J. D.: GeneMark: parallel gene recognition for both DNA strands. *Comput Chem*, 17:123–153, (1993).

Boser B. E., Guyon I. M., Vapnik V.: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, chap. A training algorithm for optimal margin classifiers, pages 144–152. ACM Press, Pittsburgh, PA (1992).

Breitbart M., Felts B., Kelley S., Mahaffy J. M., Nulton J., Salamon P., Rohwer F.: Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci*, 271:565–574, (2004a).

Breitbart M., Hewson I., Felts B., Mahaffy J. M., Nulton J., Salamon P., Rohwer F.: Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol*, 185:6220–6223, (2003).

Breitbart M., Miyake J. H., Rohwer F.: Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol Lett*, 236:249–256, (2004b).

Breitbart M., Salamon P., Andresen B., Mahaffy J. M., Segall A. M., Mead D., Azam F., Rohwer F.: Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*, 99:14250–14255, (2002).

Béjà O., Aravind L., Koonin E. V., Suzuki M. T., Hadd A., Nguyen L. P., Jovanovich S. B., Gates C. M., Feldman R. A., Spudich J. L., Spudich E. N., DeLong E. F.: Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289:1902–1906, (2000).

Béjà O., Suzuki M. T., Heidelberg J. F., Nelson W. C., Preston C. M., Hamada T., Eisen J. A., Fraser C. M., DeLong E. F.: Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature*, 415:630–633, (2002).

Cai Y. D., Lin S. L.: Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim Biophys Acta*, 30:127–33, (2003).

Chan E. Y.: Advances in sequencing technology. *Mutat Res*, 573:13–40, (2005).

Chothia C., Gough J., Vogel C., Teichmann S. A.: Evolution of the protein repertoire. *Science*, 300:1701–1703, (2003).

Dear P. H.: One by one: Single molecule tools for genomics. *Brief Funct Genomic Proteomic*, 1:397–416, (2003).

Delcher A. L., Bratke K. A., Powers E. C., Salzberg S. L.: Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23:673–679, (2007).

Delcher A. L., Harmon D., Kasif S., White O., Salzberg S. L.: Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*, 27:4636–4641, (1999).

DeLong E. F.: Microbial community genomics in the ocean. *Nat Rev Microbiol*, 3:459–469, (2005).

Durbin R., Eddy S., Krogh A., Mitchinson G.: *Biological sequence analysis*. Cambridge University Press, Cambridge, UK (1998).

Edwards R. A., Rodriguez-Brito B., Wegley L., Haynes M., Breitbart M., Peterson D., Saar M., Alexander S., Alexander E. C., Rohwer F.: Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics*, 7:57, (2006).

Edwards R. A., Rohwer F.: Viral metagenomics. *Nat Rev Microbiol*, 3:504–510, (2005).

Eisen J. A.: Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes. *PLoS Biol*, 5:e82, (2007).

Fauth J. E., Bernardo J., Camara M., Resetarits W. J., Buskirk J. V., McCollum S. A.: Simplifying the jargon of community ecology: A conceptual approach. *The American Naturalist*, 147:282–286, (1996).

Felsenstein J.: *Inferring phylogenies*. Sinauer Associates, Sunderland, MA (2004).

Field D., Morrison N., Selengut J., Sterks P.: eGenomics: Cataloguing Our Complete Genome Collection II. *OMICS*, 10:100–1004, (2006).

Field D., Tiwari B., Snape J.: Bioinformatics and data management support for environmental genomics. *PLoS Biol*, 3:1001–1002, (2005).

Finn R. D., Mistry J., Schuster-Böckler B., Griffiths-Jones S., Hollich V., Lassmann T., Moxon S., Marshall M., Khanna A., Durbin R., Eddy S. R., Sonnhammer E. L. L., Bateman A.: Pfam: clans, web tools and services. *Nucleic Acids Res*, 34:D247–D251, (2006).

Fitch W. M.: Homology a personal view on some of the problems. *Trends Genet*, 16:227–231, (2000).

Fleischmann R. D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R., Bult C. J., Tomb J. F., Dougherty B. A., Merrick J. M.: Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269:496–512, (1995).

Fowler J., Cohen L., Jarvis P.: *Practical Statistics for Field Biology*. Wiley, New York, YK (1998).

Fraser C. M., Eisen J., Fleischmann R. D., Ketchum K. A., Peterson S.: Comparative genomics and understanding of microbial biology. *Emerg Infect Dis*, 6:505–512, (2000a).

Fraser C. M., Eisen J. A., Nelson K. E., Paulsen I. T., Salzberg S. L.: The value of complete microbial genome sequencing (you get what you pay for). *J Bacteriol*, 184:6403–5; discusion 6405, (2002).

Fraser C. M., Eisen J. A., Salzberg S. L.: Microbial genome sequencing. *Nature*, 406:799–803, (2000b).

Fraser C. M., Fleischmann R. D.: Strategies for whole microbial genome sequencing and analysis. *Electrophoresis*, 18:1207–1216, (1997).

Fraser C. M., Norris S. J., Weinstock G. M., White O., Sutton G. G., Dodson R., Gwinn M., Hickey E. K., Clayton R., Ketchum K. A., Sodergren E., Hardham J. M., McLeod M. P., Salzberg S., Peterson J., Khalak H., Richardson D., Howell J. K., Chidambaram M., Utterback T., McDonald L., Artiach P., Bowman C., Cotton M. D., Fujii C., Garland S., Hatch B., Horst K., Roberts K., Sandusky M., Weidman J., Smith H. O., Venter J. C.: Complete genome sequence of Treponema pallidum, the syphilis spirochete. *Science*, 281:375–388, (1998).

Frishman D., Mironov A., Mewes H., Gelfand M.: Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res*, 26:2941–2947, (1998).

Furey T. S., Cristianini N., Duffy N., Bednarski D. W., Schummer M., Haussler D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, (2000).

Furrie E.: A molecular revolution in the study of intestinal microflora. *Gut*, 55:141–143, (2006).

Gao H., Yang Z. K., Gentry T. J., Wu L., Schadt C. W., Zhou J.: Microarray-based analysis of microbial community RNAs by whole-community RNA amplification. *Appl Environ Microbiol*, 73:563–571, (2007).

Gentry T. J., Wickham G. S., Schadt C. W., He Z., Zhou J.: Microarray applications in microbial ecology research. *Microb Ecol*, 52:159–175, (2006).

Gill S. R., Pop M., Deboy R. T., Eckburg P. B., Turnbaugh P. J., Samuel B. S., Gordon J. I., Relman D. A., Fraser-Liggett C. M., Nelson K. E.: Metagenomic analysis of the human distal gut microbiome. *Science*, 312:1355–1359, (2006).

Gould S. J.: *Full house: The spread of excellence from Plato to Darwin.*, page 244. Harmony Books, New York, NY (1996).

Gouy M., Gautier C.: Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res*, 10:7055–7074, (1982).

Green R. E., Krause J., Ptak S. E., Briggs A. W., Ronan M. T., Simons J. F., Du L., Egholm M., Rothberg J. M., Paunovic M., Pääbo S.: Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444:330–336, (2006).

Guo F.-B., Ou H.-Y., Zhang C.-T.: ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res*, 31:1780–1789, (2003).

Hansen S. K., Rainey P. B., Haagensen J. A. J., Molin S.: Evolution of species interactions in a biofilm community. *Nature*, 445:533–536, (2007).

Hastie T., Tibshirani R., Friedman J.: *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY (2002).

Hou Y., Hsu W., Lee M. L., Bystroff C.: Efficient remote homology detection using local structure. *Bioinformatics*, 19:2294–2301, (2003).

Hugenholtz P.: Exploring prokaryotic diversity in the genomic era. *Genome Biol*, 3:REVIEWS0003, (2002).

Hugenholtz P., Goebel B. M., Pace N. R.: Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol*, 180:4765–4774, (1998).

Jaakkola T., Diekhans M., Haussler D.: A discriminative framework for detecting remote protein homologies. *J Comput Biol*, 7:95–114, (2000).

Koski L. B., Golding G. B.: The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 52:540–542, (2001).

Lafay B., Lloyd A. T., McLean M. J., Devine K. M., Sharp P. M., Wolfe K. H.: Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res*, 27:1642–1649, (1999).

Larsen T. S., Krogh A.: EasyGene – a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, 4:15, (2003).

Leslie C., Eskin E., Noble W. S.: The spectrum kernel: a string kernel for SVM protein classification. *Pac Symp Biocomput*, 7:564–575, (2002).

Lewin B.: *GENES VIII*. Pearson Prentice Hall, Upper Saddle River, NJ (2004).

Linke B., McHardy A., Neuweger H., Krause L., Meyer F.: REGANOR : A Gene Prediction Server for Prokaryotic Genomes and a Database of High Quality Gene Predictions for Prokaryotes. *Appl Bioinformatics*, 5:193–198, (2006).

Lorenz P., Eck J.: Metagenomics and industrial applications. *Nat Rev Microbiol*, 3:510–516, (2005).

Lowe T. M., Eddy S. R.: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25:955–964, (1997).

Lukashin A. V., Borodovsky M.: Genemark.hmm: new solutions for gene finding. *Nucleic Acids Res*, 26:1107–1115, (1998).

Mahony S., McInerney J. O., Smith T. J., Golden A.: Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models. *BMC Bioinformatics*, 5:23, (2004).

Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y.-J., Chen Z., Dewell S. B., Du L., Fierro J. M., Gomes X. V., Godwin B. C., He W., Helgesen S., Ho C. H., Irzyk G. P., Jando S. C., Alenquer M. L. I., Jarvie T. P., Jirage K. B., Kim J.-B., Knight J. R., Lanza J. R., Leamon J. H., Lefkowitz S. M., Lei M., Li J., Lohman K. L., Lu H., Makhijani V. B., McDade K. E., McKenna M. P., Myers E. W., Nickerson E., Nobile J. R., Plant R., Puc B. P., Ronan M. T., Roth G. T., Sarkis G. J., Simons J. F., Simpson J. W., Srinivasan M., Tartaro K. R., Tomasz A., Vogt K. A., Volkmer G. A., Wang S. H., Wang Y., Weiner M. P., Yu P., Begley R. F., Rothberg J. M.: Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, (2005).

Martin H. G., Ivanova N., Kunin V., Warnecke F., Barry K. W., McHardy A. C., Yeates C., He S., Salamov A. A., Szeto E., Dalin E., Putnam N. H., Shapiro H. J., Pangilinan J. L., Rigoutsos I., Kyrpides N. C., Blackall L. L., McMahon K. D., Hugenholtz P.: Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol*, 24:1263–1269, (2006).

McHardy A. C., Goesmann A., Pühler A., Meyer F.: Development of joint application strategies for two microbial gene finders. *Bioinformatics*, 20:1622–1631, (2004a).

McHardy A. C., Martín H. G., Tsirigos A., Hugenholtz P., Rigoutsos I.: Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, 4:63–72, (2007).

McHardy A. C., Pühler A., Kalinowski J., Meyer F.: Comparing expression level-dependent features in codon usage with protein abundance: An analysis of 'predictive proteomics'. *Proteomics*, 4:46–58, (2004b).

Moore J. E., Lake J. A.: Gene structure prediction in syntenic DNA segments. *Nucleic Acids Res*, 31:7271–7279, (2003).

Morin P. J.: *Community Ecology*. Blackwell Science, Malden, MA (1999).

Moszer I., Rocha E. P., Danchin A.: Codon usage and lateral gene transfer in Bacillus subtilis. *Curr Opin Microbiol*, 2:524–528, (1999).

Médigue C., Rouxel T., Vigier P., Hénaut A., Danchin A.: Evidence for horizontal gene transfer in Escherichia coli speciation. *J Mol Biol*, 222:851–856, (1991).

Nekrutenko A., Chung W. Y., Li W. H.: An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet*, 19:306–310, (2003a).

Nekrutenko A., Chung W.-Y., Li W.-H.: ETOPE: Evolutionary test of predicted exons. *Nucleic Acids Res*, 31:3564–3567, (2003b).

Nierman W. C., Eisen J. A., Fleischmann R. D., Fraser C. M.: Genome data: what do we learn? *Curr Opin Struct Biol*, 10:343–348, (2000).

Noble W. S.: *Kernel Methods in Computational Biology*, chap. Support vector machine applications in computational biology, pages 71–92. MIT Press, Cambridge, MA (2004).

Noguchi H., Park J., Takagi T.: MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*, 34:5623–5630, (2006).

Noonan J. P., Coop G., Kudaravalli S., Smith D., Krause J., Alessi J., Chen F., Platt D., Pääbo S., Pritchard J. K., Rubin E. M.: Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314(5802):1113–1118, (2006).

Obenauer J. C., Denson J., Mehta P. K., Su X., Mukatira S., Finkelstein D. B., Xu X., Wang J., Ma J., Fan Y., Rakestraw K. M., Webster R. G., Hoffmann E., Krauss S., Zheng J., Zhang Z., Naeve C. W.: Large-scale sequence analysis of avian influenza isolates. *Science*, 311:1576–1580, (2006).

Osterman A., Overbeek R.: Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol*, 7:238–251, (2003).

Pace N. R., Stahl D. A., Lane D., Olsen G. J.: Analyzing natural microbial populations by rRNA sequences. *ASM News*, 51:4–12, (1985).

Pennisi E.: Metagenomics. Massive microbial sequence project proposed. *Science*, 315:1781, (2007).

Poinar H. N., Schwarz C., Qi J., Shapiro B., Macphee R. D. E., Buigues B., Tikhonov A., Huson D. H., Tomsho L. P., Auch A., Rampp M., Miller W., Schuster S. C.: Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, 311:392–394, (2006).

Poole F. L., Gerwe B. A., Hopkins R. C., Schut G. J., Weinberg M. V., Jenney F. E., Adams M. W. W.: Defining genes in the genome of the hyperthermophilic archaeon Pyrococcus furiosus: implications for all microbial genomes. *J Bacteriol*, 187:7325–7332, (2005).

Ram R. J., Verberkmoes N. C., Thelen M. P., Tyson G. W., Baker B. J., Blake R. C., Shah M., Hettich R. L., Banfield J. F.: Community proteomics of a natural microbial biofilm. *Science*, 308:1915–1920, (2005).

Ramaswamy S., Tamayo P., Rifkin R., Mukherjee S., Yeang C. H., Angelo M., Ladd C., Reich M., Latulippe E., Mesirov J. P., Poggio T., Gerald W., Loda M., Lander E. S., Golub T. R.: Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 98:15149–15154, (2001).

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

Riesenfeld C. S., Schloss P. D., Handelsman J.: Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*, 38:525–552, (2004).

Rusch D. B., Halpern A. L., Sutton G., Heidelberg K. B., Williamson S., Yooseph S., Wu D., Eisen J. A., Hoffman J. M., Remington K., Beeson K., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews-Pfannkoch C., Venter J. E., Li K., Kravitz S., Heidelberg J. F., Utterback T., Rogers Y.-H., Falcón L. I., Souza V., Bonilla-Rosso G., Eguiarte L. E., Karl D. M., Sathyendranath S., Platt T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari M. R., Strausberg R. L., Nealson K., Friedman R., Frazier M., Venter J. C.: The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol*, 5:e77, (2007).

Saitou N., Nei M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4:406–425, (1987).

Salzberg S. L., Delcher A. L., Kasif S., White O.: Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*, 26:544–548, (1998).

Sanger F., Air G. M., Barrell B. G., Brown N. L., Coulson A. R., Fiddes C. A., Hutchison C. A., Slocombe P. M., Smith M.: Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265:687–695, (1977a).

Sanger F., Nicklen S., Coulson A. R.: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74:5463–5467, (1977b).

Schloss P. D., Handelsman J.: Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol*, 6:229, (2005).

Schölkopf B., Smola A. J.: *Learning with Kernels*. The MIT Press, Cambridge, MA (2002).

Shannon C., Weaver W.: *The mathematical theory of communication*. Urbana, Univerisity of Illinois Press, Urbana, IL (1963).

Shibuya T., Rigoutsos I.: Dictionary-driven prokaryotic gene finding. *Nucleic Acids Res*, 30:2710–2725, (2002).

Shigenobu S., Watanabe H., Hattori M., Sakaki Y., Ishikawa H.: Genome sequence of the endocellular bacteria symbiont of aphids *Buchnera* sp. APS. *Nature*, 407:81–86, (2000).

Shmatkov A. M., Melikyan A. A., Chernousko F. L., Borodovsky M.: Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. *Bioinformatics*, 15:874–886, (1999).

Skovgaard M., Jensen L. J., Brunak S., Ussery D., Krogh A.: On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*, 17:425–428, (2001).

Smith M. W., Feng D. F., Doolittle R. F.: Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem Sci*, 17:489–493, (1992).

Spencer D. H., Kas A., Smith E. E., Raymond C. K., Sims E. H., Hastings M., Burns J. L., Kaul R., Olson M. V.: Whole-genome sequence variation among multiple isolates of Pseudomonas aeruginosa. *J Bacteriol*, 185:1316–1325, (2003).

Sterky F., Lundeberg J.: Sequence analysis of genes and genomes. *J Biotechnol*, 76:1–31, (2000).

Studier J., Keppler K.: A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5:729–731, (1988).

Tech M., Merkel R.: YACOP: Enhanced Gene Prediction Obtained by a Combination of Existing Methods. *In Silico Biol*, 3:441–51, (2003).

Teeling H., Meyerdierks A., Bauer M., Amann R., Glöckner F. O.: Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol*, 6:938–947, (2004).

Thompson J. R., Pacocha S., Pharino C., Klepac-Ceraj V., Hunt D. E., Benoit J., Sarma-Rupavtarm R., Distel D. L., Polz M. F.: Genotypic diversity within a natural coastal bacterioplankton population. *Science*, 307:1311–1313, (2005).

Tringe S. G., Rubin E. M.: Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet*, 6:805–814, (2005).

Tringe S. G., von Mering C., Kobayashi A., Salamov A. A., Chen K., Chang H. W., Podar M., Short J. M., Mathur E. J., Detter J. C., Bork P., Hugenholtz P., Rubin E. M.: Comparative metagenomics of microbial communities. *Science*, 308:554–557, (2005).

Turnbaugh P. J., Ley R. E., Mahowald M. A., Magrini V., Mardis E. R., Gordon J. I.: An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444:1027–1031, (2006).

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

Vapnik V.: *The Nature of Statistical Learning Theory*. Springer, New York, NY (1995).

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

Whitham T. G., Bailey J. K., Schweitzer J. A., Shuster S. M., Bangert R. K., LeRoy C. J., Lonsdorf E. V., Allan G. J., DiFazio S. P., Potts B. M., Fischer D. G., Gehring C. A., Lindroth R. L., Marks J. C., Hart S. C., Wimp G. M., Wooley S. C.: A framework for community and ecosystem genetics: from genes to ecosystems. *Nat Rev Genet*, 7:510–523, (2006).

Woese C. R.: Bacterial evolution. *Microbiol Rev*, 51:221–271, (1987).

Woese C. R., Fox G. E.: Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74:5088–5090, (1977).

Zuckerkandl E., Pauling L.: Molecules as documents of evolutionary history. *J Theor Biol*, 8:357–366, (1965).

CHAPTER 6

Papers

Paper I

540–549 *Nucleic Acids Research, 2007, Vol. 35, No. 2* *Published online 14 December 2006*
doi:10.1093/nar/gkl1083

GISMO—gene identification using a support vector machine for ORF classification

Lutz Krause*, Alice C. McHardy[1], Tim W. Nattkemper, Alfred Pühler, Jens Stoye and Folker Meyer[2]

Center for Biotechnology, Bielefeld University (CeBiTec), D-33594 Bielefeld, Germany, [1]Bioinformatics and Pattern Discovery Group, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA and [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received September 4, 2006; Revised November 22, 2006; Accepted November 24, 2006

ABSTRACT

We present the novel prokaryotic gene finder GISMO, which combines searches for protein family domains with composition-based classification based on a support vector machine. GISMO is highly accurate; exhibiting high sensitivity and specificity in gene identification. We found that it performs well for complete prokaryotic chromosomes, irrespective of their GC content, and also for plasmids as short as 10 kb, short genes and for genes with atypical sequence composition. Using GISMO, we found several thousand new predictions for the published genomes that are supported by extrinsic evidence, which strongly suggest that these are very likely biologically active genes. The source code for GISMO is freely available under the GPL license.

INTRODUCTION

Since the mid-1990s, automated gene finders for prokaryotic genome sequences have become available that allow the unsupervised discovery of genes from raw genomic sequence (1–9). This accomplishment, accompanied by impressive values of accuracy, has made prokaryotic gene prediction one of the showcases of computational biology. Subsequent developments have focused mostly on the introduction of novel techniques to more accurately capture sequence composition (4), modeling of the gene structure (7,10) and development of models that allow the unsupervised discovery of multiple gene classes (8,11).

Because of the high accuracy initially reported for most programs, some might consider prokaryotic gene prediction solved, but from the point of a practitioner, this is not quite the case yet. For some programs the predictive accuracy is uncertain, as they have not been re-evaluated since the original evaluation on a handful of genomes. The recent development of techniques that improve predictions by combining the output of multiple programs (6,12,13) shows that accuracy can be increased. Another issue is that some programs are only accessible via a web interface, which for genome projects—due to the confidentiality of the data—is frequently not an option.

Here we describe our novel gene finder GISMO (Gene Identification using a Support Vector Machine for ORF classification), which is freely available under the GPL license. GISMO has high classification accuracy: it is very sensitive, meaning that it identifies most known genes, and specific, i.e. it produces reliable predictions. Our program combines a hidden Markov model (HMM)-based search for protein domains with a support vector machine (SVM) to identify coding regions based on sequence composition. An advantage of the HMM-based search for protein domains compared with pair-wise sequence searches is the higher accuracy in discriminating between signal and noise for protein family members (14). Also, genes with new orderings of known protein domains can be detected easily. An SVM classifier is constructed for composition-based identification of protein-encoding genes. The SVM is a machine learning technique with a strong theoretical foundation (15,16) that has been used to improve classification accuracy in biological applications such as the detection of protein family members (17–19), RNA and DNA binding proteins (20), and the functional classification of gene expression data (21). The SVM is a maximum margin classifier that can solve non-linear classification problems by learning an optimally separating hyperplane in a higher-dimensional feature space. By use of non-linear kernel functions such as a Gaussian kernel, complex and non-linear decision functions can be learned by the SVM. Even if items of one class are clustered in multiple separate sub-regions in the input space they can be clearly separated from the other class (Figure 1). The learnt hyperplane allows accurate discrimination between classes that cannot be separated linearly in the input space, as may be the case when phenomena such as horizontal

*To whom correspondence should be addressed. Tel: +49 521 106 4823; Fax: +49 521 106 6419; Email: lutz.krause@cebitec.uni-bielefeld.de

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

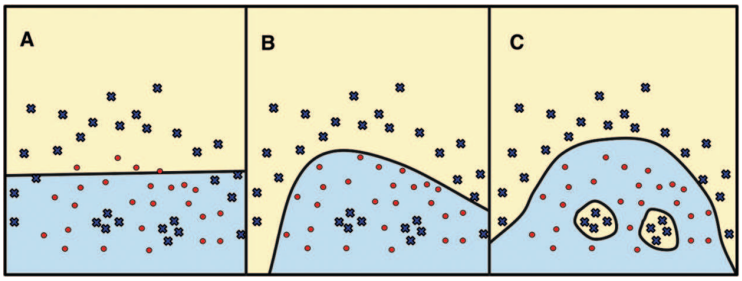


Figure 1. Class boundaries learned by the SVM with different kernel functions. Circles and crosses represent instances of a toy example training set. Colored regions indicate the two classes learned by three example SVM applications. (**A**) A linear decision function learned with a linear kernel. (**B**) A polynomial kernel allows realization of a polynomial separating surface. (**C**) With a Gaussian kernel the SVM can learn disjoint decision functions that surround a multitude of 'islands' of items from the same class (30).

gene transfer, translational selection and leading/lagging strand biases influence the sequence composition of genes (22–24).

GISMO was evaluated with 165 prokaryotic chromosomes and 223 plasmid sequences. For the chromosomal sequences, GISMO identified 94.3% of the genes (98.9% for genes with annotated function), and 94.3% of its predictions corresponded to annotated genes. Several thousand of the new predictions for the published genomes are supported by extrinsic evidence, suggesting that these very probably are biologically active genes that are missing in the annotations. We also address some of the most challenging problems for prokaryotic gene finders, including the correct identification of short genes (7,25) and of genes with atypical sequence composition and the prediction of genes when only little sequence material is available, as in the case of extrachromosomal replicons. The composition-based SVM, which uses vectors of sequence composition in the (low-dimensional) space of codon usage, is well suited for these tasks and achieved the highest classification accuracy for all cases when compared with two other popular, freely available programs. GISMO predictions for the 165 genomic sequences are available for download in GFF at http://www.CeBiTec.Uni-Bielefeld.DE/groups/brf/software/gismo.

MATERIALS AND METHODS

Datasets

The annotation and genomic sequence of 165 bacterial and archaeal chromosomes were downloaded from EMBL (26), and 223 plasmids longer than 10 kb were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/). Annotated genes tagged as pseudogenes or not corresponding to an open reading frame (beginning with a start codon, ending with an in-frame stop codon, no internal stop codon) were excluded from the reference set of annotated genes. Sets of function-known genes were created based on the gene product description. All genes supported by evidence, such as an annotated function or gene product, noted sequence conservation, or experimental support, were included in these sets. Short genes were defined as genes with <300 bp of sequence. Putative horizontally transferred genes were obtained for 57 genomes from HGT-DB (27), all having >100 genes predicted horizontally transferred. The sequences used in this study as well as tables with evaluation details are available at http://www.CeBiTec.Uni-Bielefeld.DE/groups/brf/software/gismo.

Gene-finding algorithm

GISMO proceeds in three phases: (i) an initial search for extrinsic support with HMM profiles of protein domains, (ii) the training and application of an SVM-based intrinsic classifier, and (iii) the merger of the different sources of evidence and prediction of optimal start sites.

In the first phase, the forward and reverse strand of the DNA sequence are translated in all three reading frames, and the translations are searched for protein domains contained in the Pfam-A database (28). Significant hits to the protein domain models (e-value <0.01) are mapped onto the open reading frames (ORFs) at the appropriate position in the genomic sequence. These ORFs constitute the initial set of domain-supported genes.

In the next phase, a composition-based SVM classifier is trained and applied for gene identification. All genes carrying a strongly supported domain motif (e-value $<10^{-40}$) are used as training instances for the CDS (coding sequence) class, and ORFs located in the 'shadow' of these genes are used as the training items for the non-coding ORF (nORF) class. More specifically, shadow ORFs are used that are located on another frame with an overlap of $\geqslant$90 bp with a domain-supported gene. As input to the SVM classifier, all ORFs are represented as vectors of sequence composition features. We evaluated 10 feature types for their suitability as input: oligonucleotides of length 3–9, amino acids and di-amino acids, and a combination of codons and amino acids.

Vectors of sequence composition features are composed from different sequence features F. In the case of oligonucleotide features each F is the list of all words of one chosen length k over the alphabet of all nucleotides {a,c,g,t} (for k = 3: F = aaa,aac,..,ttt; for k = 4: F = aaaa,aaac, . . . , tttt). For the amino acid and di-amino acid feature type each F is defined in an analogous way: Here, F is the list of all amino acids and di-amino acids, respectively (F = Ala,Arg, . . . , Val for the amino acid feature type; F = AlaAla, AlaArg, . . . , ValVal for the di-amino acid feature type). Now, let f_i be the feature at position i in one F. To represent each ORF x by a vector $v = (v_1, \ldots, v_c)$ of sequence composition features, we evaluate the frequencies of all f_i in x.

For the oligonucleotide feature type, only 'in-frame' oligonucleotides, i.e. oligonucleotides beginning at positions 1,4,7,... of x, are considered to account for the 3-periodicity of the genetic code. v_i is the in-frame frequency of oligonucleotide f_i in x, divided by the normalization factor r:

$$v_i = \frac{\text{frequency of oligonucleotide } f_i \text{ at position } 1,4,7,\ldots \text{of } x}{r}$$

The normalization factor r for an ORF of length n is $r = n/3$ for $k = 3$, $r = n/3-1$ for $k \in \{4,5,6,\}$ and $r = n/3-2$ for $k \in \{7,8,9\}$.

For the (di-) amino acid feature type, v_i are defined as:

$$v_i = \frac{\text{frequency of (di-) amino acid } f_i \text{ in the translated sequence of } x}{r}$$

where f_i is the (di-) amino acid at position i in F, and the normalization factor r for an ORF of length n is $r = n/3$ for amino acids and $r = n/3-1$ for di-amino acids.

We found that 64-dimensional vectors of relative codon frequencies (i.e. in-frame oligonucleotide trimers) allow the most accurate discrimination between genes and nORFs and are particularly well suited for the identification of short genes and the training of accurate classifiers for plasmid sequences. The SVM classifier is trained with a Gaussian kernel function (30). Therefore, all CDS and nORFs from the training set are implicitly mapped from the input space of sequence composition to the feature space determined by the Gaussian kernel. In this feature space a hyperplane is learned by the SVM that optimally separates all training ORFs from the two classes. A suitable Gaussian kernel parameter γ and SVM parameter C (see the following section 'support vector machine algorithm') are determined in a grid search of the parameter space by fivefold cross-validation on the training set: The training set is partitioned into five subsamples. In five steps one sample is retained as the validation set, and an SVM is trained on all remaining samples. In each step the validation set is classified with the trained SVM, and the achieved classification accuracy is measured. The cross-validation process is repeated in a grid search for different values for the Gaussian kernel parameter γ and for the SVM parameter C. Finally, values for γ and C that result in the best classification accuracy are chosen.

Subsequently, all ORFs longer than a specified minimum length (set by the user) are extracted from the genomic sequence and represented by a sequence composition vector. These sequence composition vectors are mapped to the feature space determined by the Gaussian kernel. A class is assigned to each vector depending on its relative location with respect to the learned separating hyperplane. Based on the distance to the learned hyperplane, an additional score, called the SVM-score, can be calculated and used to classify a novel item in one of the two classes (see the SVM algorithm below).

In the third phase, domain- and composition-supported CDSs are combined into one set. Gene starts are adjusted from the 'longest possible coding sequence' to alternative positions. All predictions supported by strong evidence for the existence of a protein domain (e-value <0.01) or characteristic CDS sequence composition (SVM-score >-0.1) are kept. CDS candidates supported by weaker evidence are removed if they overlap more than 50 bp with a reliable candidate.

SVM algorithm

The SVM (15,16) is a supervised learning algorithm with a strong theoretical foundation and high classification accuracy for many applications. SVMs can learn accurate classifiers for data sets that cannot be linearly separated in the input space (30). This is achieved by the choice of a suitable kernel function to transform the input data into another feature space where it is easier to compute an accurate classification (Figure 1). By learning the optimal separating hyperplane in this feature space, a non-linear classifier can be learned in the original input space. In the case of GISMO, each item of the training set (CDSs and nORFs) is represented by a vector $\boldsymbol{v}$ of its sequence composition features. Given a training set of m vectors $\boldsymbol{v}_j = (v_1, \ldots, v_c)_j$ $(1 \leqslant j \leqslant m)$ with known class labels $y_j \in \{+1, -1\}$(+1 for CDS, -1 for nORF), the SVM in training learns a hyperplane $(\boldsymbol{w}, b)$ that optimally separates the items of the two classes. The vector $\boldsymbol{w}$ that is learned by an SVM is defined as

$$w = \sum_{j=1}^{m} a_j y_j \boldsymbol{v}_j,$$

where a_j are weights that are assigned to each $\boldsymbol{v}_j$ during training. b is a scalar (29). With a learned hyperplane $(\boldsymbol{w}, b)$, a query vector $\boldsymbol{v}$ (an ORF represented by its vector of sequence composition) can be classified based on the decision value (the svm-score):

$$d(\boldsymbol{v}) = \sum_{j=1}^{m} a_j y_j k(\boldsymbol{v}, \boldsymbol{v}_j) + b,$$

where $k(\boldsymbol{v}, \boldsymbol{v}_j)$ is a kernel function (29). In the case of GISMO, $k(\boldsymbol{v}, \boldsymbol{v}_j)$ is the Gaussian kernel: $k(\boldsymbol{v}, \boldsymbol{v}_j) = e^{-\gamma \|v\text{-}vj\|^2}$.

In other words: To calculate the decision value $d(\boldsymbol{v})$, $\boldsymbol{v}$ is compared with the sequence composition $\boldsymbol{v}_j$ of each training ORF using the Gaussian kernel function. If $\boldsymbol{v}$ is more 'similar' to the CDSs from the training set a positive score is obtained, otherwise $d(\boldsymbol{v})$ is negative. Depending on whether $d(\boldsymbol{v})$ is larger than or smaller than 0, items are usually classified into one of the two classes by the SVM. To increase the sensitivity of GISMO, a relaxed cut-off is used—an ORF is classified as CDS if $d(\boldsymbol{v}) \geqslant -0.6$; otherwise it is classified as nORF.

The weights a_j that are learned during training may be bounded by a finite value C. Therefore, the SVM parameter C influences the generalization ability of the learned classifier. If C is set to a small value, outlying training items are misclassified (29); this approach can be used to reduce overfitting in the case of small training sets. If C has a finite value, the resulting classifier is called a 'soft margin SVM' (29).

With a Gaussian kernel disjoint decision functions can be realized (30). The Gaussian kernel parameter γ influences the local behavior of the learned decision boundary. Setting a value for γ is a tradeoff between a well-fitted or more

Table 1. ROC analysis of the classification accuracy achieved with different kernel functions

Organism	Accession no.	Linear	Polynomial	Gaussian	$\Delta_{\text{best-linear}}$
E.coli O157:H7	BA000007	0.960	0.960	0.968	0.008
T.pallidum subsp. *Pallidum* str. Nichols	AE000520	0.920	0.930	0.929	0.010
C.trachomatis D/UW-3/CX	AE001273	0.976	0.982	0.987	0.009
B.aphidicola str. APS	BA000003	0.986	0.991	0.989	0.005

The $ROC_{0.1}$ measures the discriminatory power of the SVM in gene identification based on sequence composition with a linear, polynomial or Gaussian kernel function.

generalized decision boundary. A large value for γ results in irregular and noisy decision boundaries that are well fit to the training data with more disjoint clusters. A small value for γ, on the other hand, results in smooth and stable boundaries that avoid overfitting and are more robust (30). With a polynomial kernel a polynomial separating surface is learned in the input space (Figure 1).

Measures of accuracy

By comparing the predicted genes with the annotated genes, one can determine the number of correct gene predictions (tp), the number of false gene predictions (fp), the number of genes that were not found (fn), and the number of correctly classified nORFs (tn).

Classification accuracy is measured by the sensitivity $Sn = \frac{tp}{tp+fn}$ (percentage of correctly identified genes) and specificity $Sp = \frac{tp}{tp+fp}$ (percentage of correct predictions). The correlation coefficient $Cor = (N \cdot Sn \cdot Sp - tp)/[(N \cdot Sn - tp) \cdot (N \cdot Sp - tp)]^{1/2}$ describes the agreement of predictions and annotation with a single value in the range of $[-1,1]$, where $N = tp + fp + tn + fn$. Only predictions and annotated CDSs with >90 bp were included in the analysis. The accuracy for predicting translation start sites was not evaluated. To predict translation start sites GISMO uses GS-Finder, which already has been found to be very accurate (31).

The receiver operating characteristic (ROC) (32) was used to evaluate the suitability of different kernel functions for gene identification with the sequence composition-based SVM classifier. Calculation of the ROC allows a comparison of different methods independent of an individual threshold setting used to discriminate between items of two classes. The ROC value corresponds to the area under a curve of the sensitivity versus the false positive prediction rate [fp/(fp + tn)] across the range of threshold settings. We here use the $ROC_{0.1}$, which corresponds to the area under the ROC curve up to a false positive prediction rate of 10%.

Accuracy of different kernels for different types of genomes

The SVM for the composition-based identification of genes can be combined with a number of kernel functions that learn different types of discriminatory functions in the input space of sequence composition (Figure 1). We evaluated the classification accuracy achievable with different kernel functions for the genomes of four organisms in detail. For each of these organisms, different properties are most pronounced in genomic sequence composition. The 5.5 Mb genome of *Escherichia coli* O157:H7 contains a 1.4 Mb large O157:H7-specific region, which has mostly been acquired by lateral transfer (33). Half of this region corresponds to 24 prophages and prophage-like elements. The codon usage of *E.coli* is also influenced by translational selection. In the space of codon usage, *E.coli* genes separate into three classes, containing horizontally acquired, typical, or highly expressed genes (23). The genome of *Treponema pallidum* has a strong strand-specific bias in codon usage, shows little evidence of translation selection (22), and contains 76 horizontally transferred genes according to HGT-DB. The codon usage of *Chlamydia trachomatis* genes reflects a complex mixture of influences, the strongest being leading/lagging strand differences and translational selection (34). Its pronounced synteny to the *C. pneumonia* genome is considered evidence of a minimal foreign gene uptake (35). The codon usage of *Buchnera aphidicola* is generally very uniform, although a slight leading/lagging strand bias is detectable (36). *B.aphidicola* is a close relative of *E.coli* but has a reduced genome that contains only a subset of 564 of the *E.coli* genes (37). It does not contain horizontally acquired genes, according to HGT-DB.

For all genomes the highest classification accuracy is achieved with one of the non-linear kernel functions (Table 1). Only for *E.coli* O157:H7 are the $ROC_{0.1}$ values obtained with the linear and polynomial kernel the same. The most accurate classification for *E.coli* O157:H7 is achieved with the Gaussian kernel. This shows that the Gaussian kernel is well suited for gene prediction in genomes with distinct gene classes. For the prediction of the genes most strongly influenced by the leading/lagging strand bias of the *T.pallidum* genome, both the polynomial and the Gaussian kernel allow a more accurate prediction. Even for the very homogeneous *B.aphidicola* genome where there is little variation in codon usage, the classification accuracy improves with the non-linear kernels. These results show how a non-linear model for genes in the codon usage space can improve the classification accuracy compared to a linear classifier.

Overall the obtained differences in ROC values between linear and non-linear kernels are low. Yet, owing to the high number of ORFs, even small differences in ROC values may indicate a considerable change in accuracy. ROC values can be interpreted as the probability that when randomly picking one positive and one negative item, the classifier will assign a higher score to the positive item than to the negative. In the optimal case, with a ROC value of 1, in 100% of cases a higher score will be assigned to the positive item. For an average genome with 3 Mb and ~3000 ORFs, a change in ROC of, say, 0.01 reflects a difference of 30 ORFs that are correctly classified.

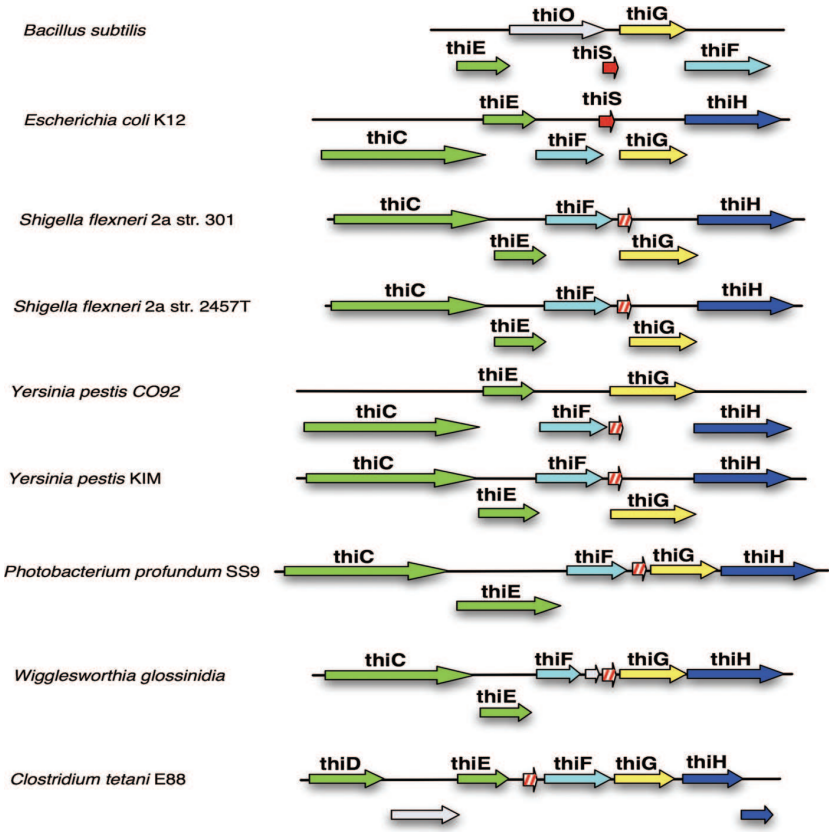


Figure 2. New candidates for the *thiS* gene of thiamin biosynthesis in the genomes of seven organisms. Homologous genes of the different organisms are drawn in the same color. The newly predicted *thiS* genes (hatched arrows) and homologs of the adjacent genes occur in conserved gene clusters. Genes with no sequence similarity to any of the displayed genes are colored grey. Overlapping genes are drawn below the continuous line. The displayed annotation is part of the thiamin biosynthesis pathway annotation of the SEED system, which is maintained and manually curated by human experts (36).

Validating novel genes by comparative analysis

To validate the novel genes predicted by GISMO for the published chromosomes, we tested for conservation of sequence, location and functional context, which is one of the strongest indicators of a biologically valid prediction (38). The chromosomal arrangement of the predictions and their surrounding genes was compared to the arrangement of homologous genes in other microbial genomes. Information about gene clusters in different organisms was obtained from the SEED, which is a manually curated comparative genomic database (39). Novel predictions homologous to non-hypothetical entries in this database (80% of the query sequence aligned using BLAST), and located in gene clusters conserved between two or more organisms (with at least two of the three adjacent upstream and downstream genes also found in the neighborhood of the homolog) were categorized as 'probably coding' (Figure 2).

Implementation

GISMO is implemented in Perl using an object-oriented approach. From the HMMER package (40), *hmmpfam* is used to search the Pfam-A database. The parallel execution of *hmmpfam* on a high performance computing resource is facilitated by the use of a DRMAA-compliant interface (http://www.drmaa.org/). The results of the domain searches with *hmmpfam* are parsed with the BioPerl library (41). For the SVM-based classification the LIBSVM library and python scripts are utilized (http://www.csie.ntu.edu.tw/~cjlin/libsvm), which perform the scaling of the input data, cross validation for model selection, training of the SVM and the SVM-based classification. The GS-Finder software is used to identify translation start sites.

RESULTS

Accuracy for prokaryotic chromosomes

The accuracy of GISMO was evaluated with 165 publicly available complete prokaryotic chromosomes. Overall, GISMO is both highly sensitive and specific in predicting prokaryotic genes, with a value of 94.3% for both measurements. For the function-known genes, which are annotated with either a functional description or experimental evidence, the sensitivity of GISMO is even 98.9%. We also found that 4336 (16.4%) of the novel GISMO predictions that are not

contained in the annotations are supported either by a significant protein domain motif (2423) or by the presence of homologs and a conserved genomic context found in the genomes of other organisms (4329, see 'Identification of novel genes in the published genomes' below).

Compared to two other popular gene finders that are freely available, GISMO is the most accurate (Table 2). Figure 3 shows a Venn diagram of the sets of genes predicted by GISMO and the gene finders Glimmer and CRITICA. We point out the high specificity of our program, which predicts ~26 000 additional genes for the 165 chromosomes in addition to the currently annotated ones, compared with ~115 000 additional predictions for Glimmer. Compared to CRITICA, which is very specific and produces reliable assignments, GISMO is more sensitive. GISMO's gene-finding accuracy does not seem significantly affected by the genomic GC content: For 42 genomes with a GC content >56%, the sensitivity is 93.5% and the specificity 92.9% (Table 2), <1% (2%) different from the overall accuracy achieved. For the function-known genes of the 42 GC-rich genomes, the sensitivity of GISMO is not reduced (99%).

Accuracy for short genes

The knowledge of the short genes of an organism is crucial because many proteins with important cellular functions are encoded by genes with <300 bp (e.g., regulatory or ribosomal proteins). Short genes are generally more difficult to identify than longer genes because their sequence carries less information that can be evaluated for classification. Figure 4 shows a comparison of the gene-finding accuracy for GISMO, Glimmer, and CRITICA for different minimum gene lengths. The results clearly show that the classification accuracy decreases with decreasing gene length. For short genes (<300 bp) GISMO has the highest overall prediction accuracy of the three programs, with a sensitivity and specificity of 63% and 69%, respectively (Table 3). CRITICA makes the most reliable predictions but identifies only 46% of the genes. Glimmer is the most sensitive (72%), but 56% of the predictions are false. Statistics suggest that a considerable fraction of the short annotated genes might, in fact, not be genes (42,43) (also, a large fraction of short genes are annotated as 'hypothetical protein'), which makes an evaluation with more reliable gene sets especially important. For the function-known genes of the short genes, GISMO is also the most sensitive program, whereby sensitivity increases by >23% to 86.4%. GISMO thus has the highest overall classification accuracy and is the most sensitive program for detection of function-known short genes.

Table 2. Gene-finding accuracy for 165 prokaryotic chromosomes

Gene finder	Cor	Sn (%)	Sp (%)
GISMO	0.94 (0.93)	94.3 (93.5)	94.3 (92.9)
Glimmer	0.87 (0.77)	94.0 (89.8)	83.3 (70.0)
CRITICA	0.92 (0.91)	88.8 (87.1)	97.1 (96.2)

The overall agreement of annotation and predictions (Cor), the sensitivity, and the specificity for the gene finders GISMO, Glimmer and CRITICA are shown. The values in parentheses are for the subset of GC-rich genomes (GC content > 56%) in the data set.

Identification of horizontally transferred genes

Genes obtained by horizontal gene transfer can possess an unusual codon usage, base composition, and GC content (44). Therefore, it can be difficult to identify these genes based on the evaluation of intrinsic sequence properties. Generative methods such as Markov chains or hidden Markov models, which create a mean-based model of sequence composition by averaging over the sequence properties of their training collections, can have difficulties with genes that are best described by more than one distribution. This issue has been addressed by the inclusion of an additional model for the genes with 'atypical' sequence composition (8). The SVM has the convenient feature that it learns to optimally discriminate the genes from the non-coding ORFs during the training phase. In an unsupervised fashion, it discovers

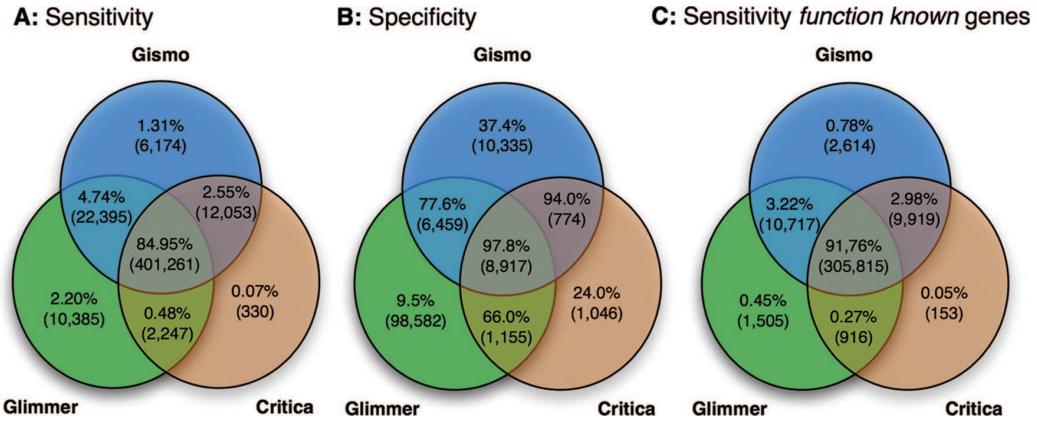


Figure 3. Comparison of the genes predicted by GISMO and two other gene finders (Glimmer and CRITICA) for 165 prokaryotic chromosomes (with 471 884 annotated genes and 333 259 function-known genes in total). (**A**) Sensitivity (percentage of identified genes) in predicting genes, whereby the numbers in the overlapping areas specify the fractions of genes identified by more than one program. The absolute numbers are given in parentheses. (**B**) Specificity (percentage of correct predictions) of predictions made by one or more of the programs. The absolute numbers of false predictions are given in parentheses. (**C**) Sensitivity for the function-known genes. The number of correct predictions is given in parentheses.

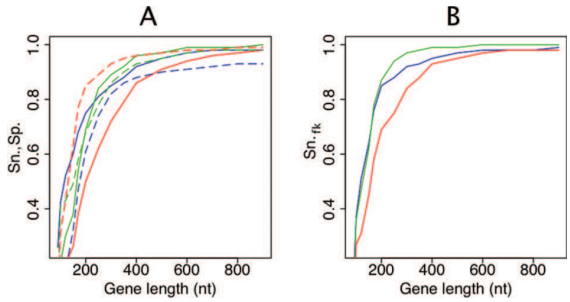


Figure 4. Sensitivity and specificity of predictions for different gene lengths. The values for GISMO, CRITICA, and Glimmer are depicted in green, red, and blue, respectively. (**A**) Relation of the gene length to the sensitivity and specificity of the three programs. The sensitivity is displayed as a solid line, the specificity as a dashed line. (**B**) Relation of the gene length to the sensitivity for function-known genes.

Table 3. Accuracy of GISMO, Glimmer and CRITICA in predicting short genes (<300 bp)

Gene finder	Cor	Sn	Sn_{fk} (%)	Sp
GISMO	0.64	63.0	86.4	69.0
Glimmer	0.54	72.0	83.7	44.0
CRITICA	0.60	46.0	67.4	84.0

Sn_{fk} denotes the sensitivity in detecting function-known genes.

Table 4. Sensitivity in the detection of probably horizontally acquired genes with atypical sequence composition

Gene finder	Sn (%)	Sn_{fk} (%)
GISMO	91.1	98.5
Glimmer	87.3	94.4
CRITICA	72.8	86.3

Sn_{fk} denotes the sensitivity in detecting function-known genes.

the shape that is most suitable for discrimination—which can be non-linear or even disjoint for genes distributed over multiple clusters in the input space (Figure 1). This property is convenient for gene prediction, as horizontal gene transfer is only one of several forces influencing sequence composition and affecting different genomes to different extents, and for each case the optimally separating boundaries can be found anew.

GISMO predicted 91.1% of the probably horizontally acquired genes with atypical sequence composition that were obtained from HGT-DB for 57 genomes. Of the function-known genes with atypical composition, 98.5% were identified, which is very similar to the overall sensitivity for prokaryotic chromosomes (98.8%), and significantly higher than the sensitivity of the other programs (Table 4).

Gene-finding accuracy for plasmids

Short DNA sequences such as plasmids yield only small sets for the training of intrinsic sequence models. For complex models with many parameters this situation can lead to overfitting and reduced prediction accuracy. GISMO is well suited for classification based on small training data sets because of (i) the use of a 'soft margin,' which allows the misclassification of outliers during training and avoids overfitting of SVMs, and (ii) the low dimensionality of the input space of codon usage, which we found to be optimal for gene prediction with a composition-based SVM.

Table 5. Gene-finding accuracy for 223 plasmids >10 kb

Gene finder	Cor	Sn (%)	Sn_{fk}(%)	Sp (%)
GISMO	0.82	89.1	96.1	80.3
Glimmer	0.79	89.3	94.4	74.5
CRITICA	0.58	45.4	55.7	87.3

Sn_{fk} denotes the sensitivity in detecting function-known genes.

For the 223 plasmid sequences >10 kb, GISMO achieves an average sensitivity of 89.1% and specificity of 80.3% and has the highest overall accuracy of the three programs (Table 5). The sensitivity increases to 96.1% for the function-known genes of the plasmids. Although this is lower than for the prokaryotic chromosomes, GISMO is very sensitive and specific, if one considers the size of the training sets available. For example, the two IncQ-like antibiotic resistance plasmids pIE1115 and pIE1130 (44) are the shortest sequences used in this survey. Both are 10 687 bp long, but differ in sequence and gene content. For pIE1115, the positive training set for the SVM consisted of five domain-supported genes, the negative training set of 60 shadow ORFs. Eight of the ten annotated genes were correctly identified, with only three additional predictions. For pIE1130, seven annotated genes were initially identified by their protein domain motifs. The training set for the composition-based classifier consisted of the seven domain-supported genes and 79 shadow ORFs. The SVM then identified two of the remaining four annotated genes, with only one additional prediction. That the classifier is able to accurately distinguish between genes and nORFs is demonstrated by the following numbers: For pIE1115, 120 of the 123 non-coding ORFs longer than 90 bp were correctly assigned, and 125 of 126 for pIE1130.

Comparison with EasyGene and GenemarkS

GISMO was also compared with the HMM-based programs EasyGene and GenemarkS, considered among the most accurate bacterial gene finders (3,7). The accuracy was evaluated on a restricted test set as both programs are only accessible via a public web interface. While GISMO and GenemarkS automatically derive training sets with genome-specific compositional sequence properties, EasyGene can be run only via its Web interface with pretrained models. Pretrained models are available only for a limited number of sequenced genomes. Therefore, the performance of GISMO, EasyGene and GenemarkS was compared for the 25 of the 365 genomic sequences for which a pretrained EasyGene model was available. For these 25 genomic sequences, all three programs display a high accuracy (Table 6). GISMO has the highest overall accuracy, with an average sensitivity and specificity of ~95%. GenemarkS has the highest average sensitivity for all genes, whereas EasyGene is most reliable. For the function-known genes, both GenemarkS and GISMO identify ~99% and thus are 3.4% more sensitive than EasyGene. EasyGene is more specific (+2.1%) but less sensitive than GISMO (−4.0%).

Table 6. Gene-finding accuracy for 25 genomic sequences

Gene finder	Cor	Sn (%)	Sn_{fk}(%)	Sp (%)
GISMO	0.943	95.1	99.0	94.7
EasyGene	0.930	91.1	95.5	96.8
GenemarkS	0.938	96.0	99.1	93.0

Sn_{fk} denotes the sensitivity in detecting function-known genes.

Identification of novel genes in the published genomes

Since the current annotations are missing many important genes (38), the novel predictions of GISMO were further investigated. For the 165 genomes used in this survey, 26 454 of the 468 368 GISMO predictions did not match an annotated gene. A strong indicator for a biologically active gene is the presence of a significant motif of a Pfam protein domain. Of the newly predicted genes, 2423 (9.2%) exhibit such motifs with strong statistical support (*E*-value $<10^{-10}$) and do not overlap with any annotated gene by >10 amino acids. An additional 4329 (16.4%) new predictions are part of conserved gene clusters found in the same or similar orders in other microbial genomes (see Section 'Material and Methods' above). In total, 4336 (16.4%) of the novel GISMO predictions are supported by external sources of evidence (Pfam hit or a conserved cluster) that suggest that these predictions are truly biologically active genes. We describe several interesting examples below:

The *thiS* gene encodes a sulfur-carrying protein that is involved in the biosynthesis of thiamin (vitamin B1) (45). The *thiS* gene has been identified in a cluster with the *thiE, thiG* and *thiF* genes in a wide range of genomes (46) but is currently unknown for *Clostridium tetani* E88, *Photobacterium profundum* SS9, *Shigella flexneri* 2a 301, *Shigella flexneri* 2a 2457T, *Wigglesworthia glossinidia, Yersinia pestis* CO92, and *Yersinia pestis* KIM. For each of these genomes, GISMO predicted a novel probably-coding gene with significant homology to known *thiS* orthologs. The novel predictions are strongly supported by their genomic context, which comprises clusters of known genes of thiamin biosynthesis (Figure 2).

GISMO also predicted 99 genes encoding ribosomal proteins that are currently missing from the genome annotations. For example, two novel GISMO predictions for *E.coli* CFT073 and *Wolinella succinogenes* DSM 1740 are very similar to the ribosomal protein L32 in *E.coli* K12 and *Helicobacter pylori* 26695. The homologs of the two novel predictions and their adjacent genes appear in a conserved order in various organisms (Figure 5). Many of the probably-coding genes are as short as the ribosomal protein-encoding genes, a situation that explains why they were missed before. In summary, our results indicate that a considerable percentage of our additional predictions are novel and currently unknown protein-encoding genes that are missing from the annotations.

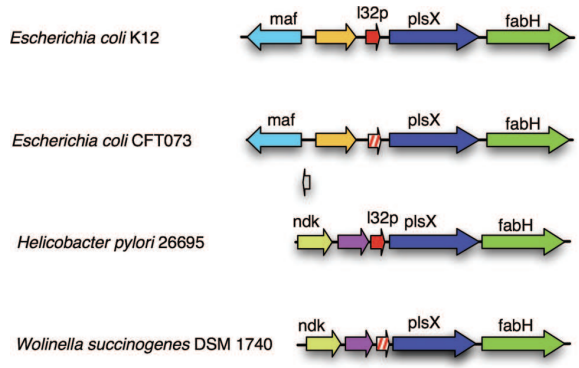


Figure 5. Candidates for missing *l*32 genes in *E.coli* CFT073 and *W.Succinogenes* DSM. Homologous genes are displayed in the same color. The novel predictions (hatched arrows) and homologs of the surrounding genes occur in conserved order in closely related genomes.

CONCLUSIONS

The gene finder GISMO presented in this work uses state-of-the-art techniques from computational biology and machine learning to accurately predict protein-encoding genes for prokaryotic genome sequences. Initially, evidence for genes is compiled by protein-domain searches with profile HMMs, which are a well-known and highly accurate means for finding members of protein families and allow a more accurate discrimination between signal and noise than pairwise sequence comparisons. They also allow the detection of genes that have protein domains in different order from that of known proteins. An SVM-based classifier is used for gene prediction based on sequence composition. The SVM is a machine learning technique that is well suited for prokaryotic gene prediction because it guarantees the unsupervised discovery of the shape in the space of sequence composition that is best suited for discrimination between genes and non-coding ORFs. The distribution of microbial genes in the space of sequence composition is affected by various influences that are pronounced to different extent for different genomes and thus require a careful and time-consuming analysis (34,36). The SVM allows the program to learn an accurate classifier, even when the distribution of items in this space is influenced by various factors such as gene expression rate, acquisition by horizontal transfer, or leading/lagging strand-related features. Gene identification for genomes in all cases was improved with non-linear classification functions, demonstrating the suitability of this approach.

In our extensive evaluation, we found GISMO to be very accurate. For the prokaryotic chromosomes, GISMO has an overall sensitivity and specificity of 94.3%. For the genes annotated with either a function or experimental evidence, the sensitivity is 98.9%. In comparison with the two popular programs Glimmer and CRITICA, which are freely available for a local installation, we found GISMO to be the most accurate also for finding genes shorter than 300 bp, for identifying genes with atypical sequence composition, and for predicting genes for short genomic sequences such as plasmid sequences. What makes this observation even more significant is the fact that GISMO is the only one of the three programs that was not used for annotating any of the genomes used in the evaluation.

In a comparison of GISMO to EasyGene and GeneMarkS on 25 genomic sequences, all three programs were very accurate, but GISMO slightly outperformed the other two in terms of overall accuracy. Therefore, GISMO presents an

open source alternative to these programs for local use and integration into genome annotation pipelines.

For the prediction of translation initiation sites, GISMO uses the GS-Finder software. Since GS-Finder has already been shown to be very accurate (31), the accuracy of GISMO in gene start site prediction was not evaluated in this survey.

For the public genomes, we found several thousand new predictions that are strongly supported by external evidence and very likely correspond to real but unannotated genes. Many of these are short, such as 99 missing ribosomal protein-encoding genes, a fact that might explain why they were not found before.

The low-dimensional input space of codon frequencies that we found to be optimal for gene identification with the SVM-based compositional classifier allows accurate classification for short genes, as well as for short genomic sequences with a low number of available training items. SVMs are also intrinsically well suited for small data sets because they avoid overfitting the learned model by using a 'soft margin' in the model optimization step.

GISMO has already been used to predict genes in more than 20 genome annotation projects, for the reannotation of genomes as well as in the international effort to annotate a thousand genomes (39). We hope that our new gene finder will be widely used in microbial genome annotation and reannotation projects and will contribute to the generation of high-quality annotations.

ACKNOWLEDGEMENTS

The authors thank Michael Dondrup, Ross Overbeek, Gordon Pusch and Gail Pieper for valuable discussions and comments. The authors also thank Santi Garcia-Vallve for supplying information for the genomes contained in HGT-DB. L.K. was supported by the DFG Graduiertenkolleg 635 Bioinformatik. F.M. was supported in part by the U.S. Department of Energy, under Contract W-31-109-Eng-38. Funding to pay the Open Access publication charges for this article was provided by the International NRW Graduate School in Bioinformatics and Genome Research.

Conflict of interest statement. None declared.

REFERENCES

1. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*., **26**, 544–548.
2. Badger,J.H. and Olsen,G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol*., **16**, 512–524.
3. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*, **29**, 2607–2618.
4. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*., **27**, 4636–4641.
5. Frishman,D., Mironov,A., Mewes,H.W. and Gelfand,M. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res*., **26**, 2941–2947.
6. Guo,F.B., Ou,H.Y. and Zhang,C.T. (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res*., **31**, 1780–1789.
7. Larsen,T.S. and Krogh,A. (2003) EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics*, **4**, 21.
8. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*., **26**, 1107–1115.
9. Shibuya,T. and Rigoutsos,I. (2002) Dictionary-driven prokaryotic gene finding. *Nucleic Acids Res*., **30**, 2710–2725.
10. Lomsadze,A., Ter-Hovhannisyan,V., Chernoff,Y.O. and Borodovsky,M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*., **33**, 6494–6506.
11. Mahony,S., McInerney,J.O., Smith,T.J. and Golden,A. (2004) Gene prediction using the self-organizing map: automatic generation of multiple gene models. *BMC Bioinformatics*, **5**, 23.
12. McHardy,A.C., Goesmann,A., Puhler,A. and Meyer,F. (2004) Development of joint application strategies for two microbial gene finders. *Bioinformatics*, **20**, 1622–1631.
13. Tech,M. and Merkl,R. (2003) YACOP: Enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol*., **3**, 441–451.
14. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol*., **284**, 1201–1210.
15. Boser,B., Guyon,I. and Vapnik,V.N. (1992) In Haussler,D. (ed.), In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press, pp. 144–152.
16. Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer.
17. Hou,Y., Hsu,W., Lee,M.L. and Bystroff,C. (2003) Efficient remote homology detection using local structure. *Bioinformatics*, **19**, 2294–2301.
18. Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol*., **7**, 95–114.
19. Leslie,C., Eskin,E. and Noble,W.S. (2002) The spectrum kernel: a string kernel for SVM protein classification. *Pac. Symp. Biocomput*., 564–575.
20. Cai,Y.D. and Lin,S.L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta*, **1648**, 127–133.
21. Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M., Jr and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
22. Lafay,B., Lloyd,A.T., McLean,M.J., Devine,K.M., Sharp,P.M. and Wolfe,K.H. (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res*., **27**, 1642–1649.
23. Medigue,C., Rouxel,T., Vigier,P., Henaut,A. and Danchin,A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol*., **222**, 851–856.
24. Moszer,I., Rocha,E.P. and Danchin,A. (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol*., **2**, 524–528.
25. Linke,B., McHardy,A.C., Neuweger,H., Krause,L. and Meyer,F. (2006) REGANOR: a gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes. *Appl. Bioinformatics*, **5**, 193–198.
26. Cochrane,G., Aldebert,P., Althorpe,N., Andersson,M., Baker,W., Baldwin,A., Bates,K., Bhattacharyya,S., Browne,P., van den Broek,A. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*., **34**, D10–D15.
27. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res*., **31**, 187–189.
28. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–141.
29. Schoelkopf,A. and Schmola,J. (2002) *Learning with Kernels*. MIT Press, Cambridge, MA.

30. Hastie,T., Tibshirani,R. and Friedman,J.H. (2003) *The Elements Of Statistical Learning*. Springer Verlag.
31. Ou,H.Y., Guo,F.B. and Zhang,C.T. (2004) GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int. J. Biochem. Cell. Biol.*, **36**, 535–544.
32. Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
33. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
34. Romero,H., Zavala,A. and Musto,H. (2000) Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res.*, **28**, 2084–2090.
35. Read,T.D., Brunham,R.C., Shen,C., Gill,S.R., Heidelberg,J.F., White,O., Hickey,E.K., Peterson,J., Utterback,T., Berry,K. *et al.* (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.*, **28**, 1397–1406.
36. Rispe,C., Delmotte,F., van Ham,R.C. and Moya,A. (2004) Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids. *Genome Res.*, **14**, 44–53.
37. Shigenobu,S., Watanabe,H., Hattori,M., Sakaki,Y. and Ishikawa,H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. *APS Nature*, **407**, 81–86.
38. Osterman,A. and Overbeek,R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.*, **7**, 238–251.
39. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
40. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
41. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
42. Skovgaard,M., Jensen,L.J., Brunak,S., Ussery,D. and Krogh,A. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.*, **17**, 425–428.
43. Nielsen,P. and Krogh,A. (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, **21**, 4322–4329.
44. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
45. Begley,T.P., Downs,D.M., Ealick,S.E., McLafferty,F.W., Van Loon,A.P., Taylor,S., Campobasso,N., Chiu,H.J., Kinsland,C., Reddick,J.J. *et al.* (1999) Thiamin biosynthesis in prokaryotes. *Arch. Microbiol.*, **171**, 293–300.
46. Rodionov,D.A., Vitreschak,A.G., Mironov,A.A. and Gelfand,M.S. (2002) Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J. Biol. Chem.*, **277**, 48949–48959.

Paper II

ARTICLES

nature biotechnology

Complete genome of the mutualistic, N_2-fixing grass endophyte *Azoarcus* sp. strain BH72

Andrea Krause[1], Adarsh Ramakumar[1], Daniela Bartels[2], Federico Battistoni[1], Thomas Bekel[2], Jens Boch[3], Melanie Böhm[1], Frauke Friedrich[1], Thomas Hurek[1], Lutz Krause[2], Burkhard Linke[2], Alice C McHardy[2,6], Abhijit Sarkar[1], Susanne Schneiker[2,4], Arshad Ali Syed[1], Rudolf Thauer[5], Frank-Jörg Vorhölter[2,4], Stefan Weidner[2], Alfred Pühler[2,4], Barbara Reinhold-Hurek[1], Olaf Kaiser[2,4,7] & Alexander Goesmann[2,7]

Azoarcus **sp. strain BH72, a mutualistic endophyte of rice and other grasses, is of agrobiotechnological interest because it supplies biologically fixed nitrogen to its host and colonizes plants in remarkably high numbers without eliciting disease symptoms. The complete genome sequence is 4,376,040-bp long and contains 3,992 predicted protein-coding sequences. Genome comparison with the *Azoarcus*-related soil bacterium strain EbN1 revealed a surprisingly low degree of synteny. Coding sequences involved in the synthesis of surface components potentially important for plant-microbe interactions were more closely related to those of plant-associated bacteria. Strain BH72 appears to be 'disarmed' compared to plant pathogens, having only a few enzymes that degrade plant cell walls; it lacks type III and IV secretion systems, related toxins and an N-acyl homoserine lactones–based communication system. The genome contains remarkably few mobile elements, indicating a low rate of recent gene transfer that is presumably due to adaptation to a stable, low-stress microenvironment.**

Endophytic bacteria reside within the living tissue of plants without substantively harming them. They are of high interest for agrobiotechnological applications, such as the improvement of plant growth and health, phytoremediation[1] or even as biofertilizer[2]. Supply of nitrogen derived from fixation of atmospheric N_2 by grass endophytes, such as *Gluconacetobacter diazotrophicus* and *Azoarcus* sp. strain BH72, which has been shown to occur in sugarcane[3] and Kallar grass[2], is a process of potential agronomical and ecological importance.

Although the lifestyle of these endophytes is relatively well documented, the molecular mechanisms by which they interact beneficially with plants have only been poorly elucidated. A combination of features makes *Azoarcus* sp. strain BH72 an excellent model grass-endophyte[4]. (i) It supplies nitrogen derived from N_2 fixation to its host, Kallar grass (*Leptochloa fusca* (L.) Kunth); *in planta* it is usually not culturable, but can be detected by culture-independent methods based on *nifH*-encoding nitrogenase reductase, the key enzyme for N_2 fixation[2]. (ii) It colonizes nondiseased plants in remarkably high numbers: estimates range from 10^8 cells (culturable cells per gram root dry weight (RDW) of field-grown Kallar grass[5]) to 10^{10} cells (estimated on the basis of abundance of bacterial *nifH*-mRNA in roots)[2]. (iii) It is the only cultured grass endophyte shown by molecular methods to be the most actively N_2-fixing bacterium of the natural population in roots[2]. (iv) It also colonizes the roots of rice, a cereal of global importance, in high numbers (10^9 cells per g RDW) in the laboratory, and spreads systemically into shoots[6]. Plant stress response is only very limited in a compatible, that is, well-colonized rice cultivar[7]. Notably, *Azoarcus* sp. strain BH72 is capable of endophytic N_2-fixation inside the roots of rice[8].

For a wider application in agriculture, more knowledge is required on mechanisms of interaction and host specificities. Although the genome of a related species, strain EbN1, belonging to a branch of *Azoarcus* species that typically occurs in soils and sediments but not in association with plants[9], is available[10], phenotypic differences and phylogenetic distances of 5–6% suggest they might deserve the rank of a separate genus in future[9]. The plant-associated strain BH72—like many N_2-fixing endophytes grass endophytes—has not been detected in root-free soil[11]. In this study, we present the complete genome sequence of a diazotrophic grass endophyte, *Azoarcus* sp. strain BH72, and highlight features that may contribute to knowledge of the endophytic lifestyle of these plant-beneficial bacteria, which may be instrumental in developing biotechnological applications.

RESULTS

General features of the genome and mobile elements

The *Azoarcus* sp. strain BH72 genome sequence was obtained with a whole genome shotgun approach, the assembly being validated by a complete fosmid (**Fig. 1b**) and a bacterial artificial chromosome

[1]Laboratory of General Microbiology, University of Bremen, PO Box 330440, D-28334 Bremen, Germany. [2]Center for Biotechnology (CeBiTec), Bielefeld University, PO Box 100131, D-33501 Bielefeld, Germany. [3]Institut für Genetik, Martin-Luther-Universität, Weinbergweg 10, D-06120 Halle/Saale, Germany. [4]Lehrstuhl für Genetik, Bielefeld University, PO Box 100131, D-33501 Bielefeld, Germany. [5]Max-Planck-Institute for Terrestrial Microbiology, Karl-von-Frisch-Strasse, D-35043 Marburg, Germany. [6]Present address: Bioinformatics & Pattern Discovery Group, IBM Thomas J Watson Research Center, Yorktown Heights, New York 10598, USA. [7]These authors contributed equally to the work. Correspondence should be addressed to Barbara Reinhold-Hurek (breinhold@uni-bremen.de).

Received 22 May; accepted 4 August; published online 22 October 2006; doi:10.1038/nbt1243

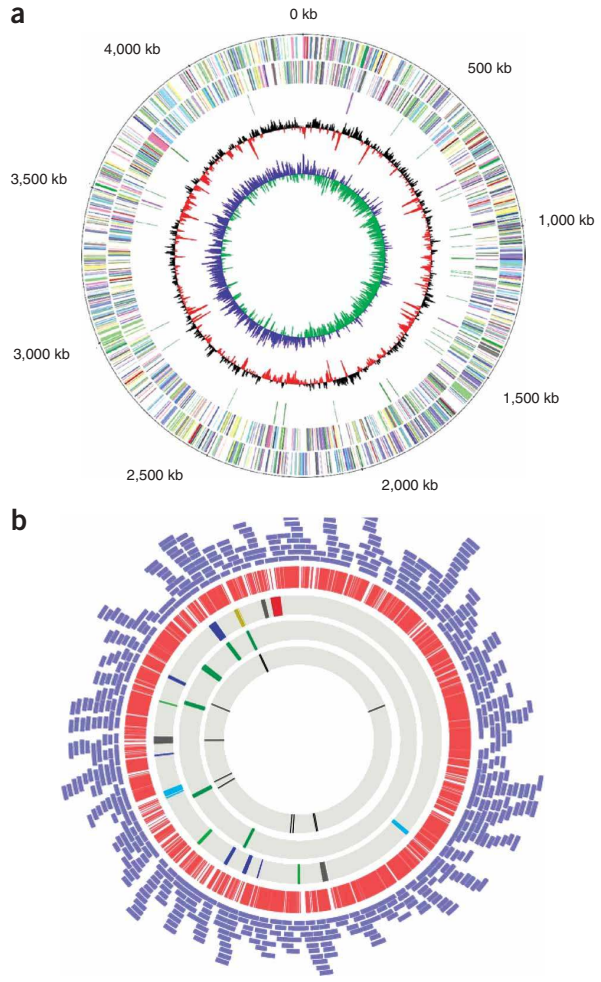


Figure 1 Circular representation of the *Azoarcus* sp. strain BH72 genome displaying relevant genome features and validation of the sequence assembly by a fosmid map. In the final consensus sequence each base matched at least phred40 quality. (**a**) From the outer to the inner concentric circle: circle 1, genomic position in kilobases; the origin of replication was clearly detectable by a bias of G toward the leading strand (GC skew); the start of the *dnaA* gene located in this region was defined as zero point of the chromosome; circles 2 and 3, predicted protein-coding sequences (CDS) on the forward (outer wheel) and the reverse (inner wheel) strands colored according to the assigned COG (clusters of orthologous groups) classes; leading strand, 2,074 CDS = 52.0%; lagging strand, 1,918 CDS = 48.0%; circle 4, tRNA genes (green) and the four rRNA operons (pink); circle 5, the G+C content showing deviations from the average (67.92%); circle 6, GC skew; a bi-directional replication mechanism suggested by a clear division into two equal replichores. (**b**) Fosmid map of the *Azoarcus* sp. strain BH72 chromosome. Each blue arc represents a single fosmid clone mapped onto the assembled sequence; circle 1, CDS with homologs in the chromosome of *Azoarcus* sp. *strain* EbN1 (e-value below e^{-30}); circle 2, gene clusters coding for surface-related proteins or other functions not related to proteins of *Azoarcus* sp. strain EbN1: exopolysaccharide/ lipopolysaccharide-related and pilus-related gene clusters (blue), flagella and chemotaxis related gene clusters (light blue), virulence-related gene clusters (red), proteins related to metabolism (gray), conserved hypothetical proteins or other proteins related to proteins of rhizobia or plant commensals, and various genes not present in *Azoarcus* sp. strain EbN1 (gold); circle 3, putative genomic islands predicted by the Pai-Ida program 1.1 (score > 3.8); circle 4, transposases and phage-related genes.

(BAC) map[12]. Characteristics of the single, circular chromosome and the predicted genes are shown in **Figure 1** and **Table 1**.

The genome contains remarkably few phage- or transposon-related genes, indicating a low degree of lateral transfer and genome rearrangements; just eight loci (**Fig. 1b**, circle 4) contain genes for integrases, recombinases, transposases or phage-related genes (**Supplementary Table 1** online). Only a few loci correspond to predicted anomalous gene clusters or putative pathogenicity islands (**Fig. 1b**, circle 3). In contrast, the genome of the *Azoarcus*-related soil isolate strain EbN1 contains 237 transposon-related genes[10]. Also rhizobial genomes harbor >100 transposases or phage-related genes (http://www.kazusa.or.jp/rhizobase/). Likewise, many plant-pathogenic proteobacteria contain high numbers of mobile elements[14]. High genomic plasticity might reflect the need for continuous adaptation to changing environments like soil or to host defense mechanisms. For nodule symbionts, soil is an alternative habitat in their life cycle; in contrast typical grass endophytes can not usually be isolated from root-free soil[11,13]. The comparatively low number of mobile elements in the endophyte BH72 might indicate a low rate of recent gene transfer and genome rearrangements, which is presumably due to adaptation to a stable, low-stress microenvironment inside plants.

Comparative genomics

Genome comparison revealed a surprisingly low degree of synteny between genomes of strain BH72 and the *Azoarcus*-related strain EbN1 (**Fig. 2**). At a low cutoff e-value of e^{-30}, the majority of predicted proteins (58%) in strain BH72 have some counterparts in strain EbN1 (**Fig. 1b**, circle 1). However, only 43% of these proteins were more closely related to those of EbN1 than to proteins of other strains. Other pathogenic or plant symbiotic proteobacteria have even less related genomes (**Supplementary Table 2** online). Because strains BH72 and EbN1 have a very different ecology, the differences may give important hints as to which genes are required specifically for the endophytic lifestyle. Several gene clusters of strain BH72 that are

Table 1 Genome features of the N_2-fixing endophyte *Azoarcus* sp. strain BH72 in comparison to the denitrifying soil bacterium *Azoarcus* sp. strain EbN1

Feature	*Azoarcus* sp. BH72	*Azoarcus* sp. EbN1
Size of chromosome (bp)	4,376,040	4,296,230
Plasmids	0	2[a]
G+C content, %	67.92	65.12
Coding sequences	3,992	4,133
Function assigned	3,418	2,560
Conserved hypothetical protein	517	628
Hypothetical protein	57	945
% of genome coding	91.2	90.9
Average length (bp)	999	945
Maximal length (bp)	6,330	6,132
% ATG initiation codons	86.57	76.46
% GTG initiation codons	10.40	16.01
% other initiation codons	n.d.[b]	n.d.
RNA elements		
rRNA operons	4	4
tRNAs	56[c]	58

[a]Plasmid 1 (207,355 bp), plasmid 2 (223,670 bp). [b]n.d., not determined. [c]One $tRNA^{Ile}$ (azo_tRNA_0051) is disrupted by a self-splicing group I intron in the CAT anticodon loop[42].

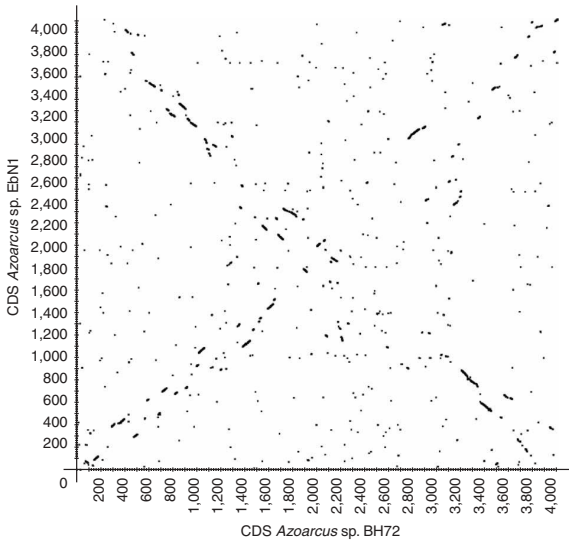


Figure 2 Synteny between the genomes of *Azoarcus* sp. strain BH72 and *Azoarcus* sp. strain EbN1. The genome of *Azoarcus* sp. strain EbN1 is adjusted to *dnaA* with its start codon as zero point of the chromosome. The diagram depicts *x-y* plots of dots forming syntenic regions between the two *Azoarcus* genomes. Each dot represents an *Azoarcus* sp. strain BH72 CDS having an ortholog in *Azoarcus* sp. strain EbN1, with coordinates corresponding to the CDS number in each genome. The orthologs were identified by best BLASTP matches of amino acid sequences (e-value $< e^{-30}$).

lacking in EbN1 or that are more similar to genes of other bacteria (**Fig. 1b**) harbor genes that encode proteins putatively involved in cell surface components or other features that may be required for the endophytic lifestyle (see below, **Fig. 3** and **Supplementary Table 1** online).

Carbon metabolism and signal transduction

Strain BH72 has a strictly respiratory type of metabolism and does not grow on any carbohydrate tested[9,15]. Aspects of putative carbon metabolism are shown in **Figure 4**. The inability to utilize common carbohydrates might contribute to a plant-compatible endophytic lifestyle because, in contrast to phytopathogens, the bacteria cannot grow and proliferate on the major cell wall constituents although a cellulase is present[16].

The major carbon sources for strain BH72 are dicarboxylic acids and ethanol[9]. Transport systems for C4-dicarboxylates (**Fig. 4**) might be of vital importance during the association with host plants, as in symbiotic rhizobia[17]. Ethanol might be important for association with flooded plants like rice, which accumulate ethanol under anoxic conditions, especially at root tips—one of the typical sites of colonization of strain BH72. Correspondingly, its genome harbors ten genes encoding putative alcohol dehydrogenases.

Strain BH72, despite being adapted to a relatively stable, low-stress microenvironment, shows a remarkable density of signal transduction systems (see details in **Supplementary Table 3** online). Thus it may be a good example for sophisticated signal transduction networks.

N_2 fixation and nitrogen metabolism

Azoarcus sp. strain BH72 appears to be highly adapted to environments poor in available nitrogen sources, which correlates with its role as an N_2-fixing endophyte (**Fig. 4**).

(i) A low-affinity glutamate dehydrogenase (GDH) for ammonium assimilation is lacking, a feature highly unusual in free-living bacteria, whereas it is present in strain EbN1. Only the high-affinity ATP-consuming assimilation system (GS[2x]-GOGAT) is present. (ii) Four genes encoding high-affinity ammonia transporters exist (*amtB/Y/D/E*), one of them with an additional regulatory domain. (iii) In contrast to the soil strain EbN1, structural genes for the molybdenum-dependent nitrogenase complex and all genes required for cofactor synthesis and maturation of the nitrogenase are present in strain BH72, one of them in two copies (*nifY*). Several putative low-potential electron donors for N_2 fixation were identified including two flavodoxin-encoding genes (*nifF1, nifF2*), 12 genes for ferredoxin-like proteins; two clusters encoding putative electron transport systems (*rnf1, rnf2*) might be instrumental for electron supply to ferredoxin during N_2 fixation. Several genes likely to be involved in the regulatory cascade are also listed in **Supplementary Table 1** online. Although in pure culture, N_2-fixing strain BH72 does not excrete substantial amounts of nitrogenous compounds[18], it supplies fixed nitrogen to its grass host[2]. The four ammonia transport proteins are putative candidates for export to the plant. Two transport systems for glutamate or glutamine as well as nine for branched-chain amino acids were also identified; however, the presence of periplasmic substrate–binding proteins suggests that these systems are used for import and not for export.

About 38 genes encoding enzymes and transporters involved in nitrate metabolism were identified (**Fig. 4** and **Supplementary Table 1** online). As in strain EbN1, genes required for assimilatory nitrate and

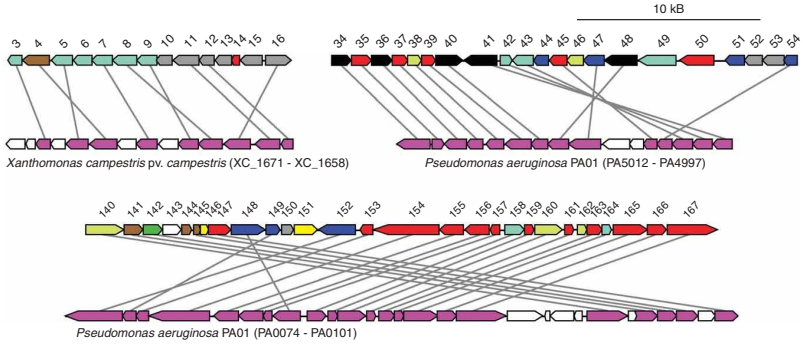


Figure 3 Gene clusters in *Azoarcus* sp. strain BH72 that are lacking in EbN1 or are more similar to genes of other bacteria. Synteny of selected clusters with gene clusters of other bacteria: *gum*-cluster, *Xanthomonas campestris* pv. *campestris* (locus_tag numbers XC_1671 - XC_1658); lipopolysaccharide-related cluster, *Pseudomonas aeruginosa* PA01 (locus_tag numbers PA5012 - PA4997); *sci*-cluster, *Pseudomonas aeruginosa* PA01 (locus_tag numbers PA0074 - PA0101); pink/white, genes present or not present in the gene cluster of strain BH72, respectively. Numbers refer to genes of *Azoarcus* sp. strain BH72 listed in **Supplementary Table 1** online. Highest similarities to proteins of other bacteria are depicted by the following colors: red, human or animal pathogens; green, root nodule symbionts (rhizobia); yellow-green, root-associated bacteria; turquoise, plant pathogens; other bacteria according to their phylogenetic affiliations: black, *Azoarcus* sp. strain EbN1; gray, beta-subgroup of *Proteobacteria*; blue, gamma-subgroup of *Proteobacteria*; yellow, alpha subgroup of *Proteobacteria*; brown, others.

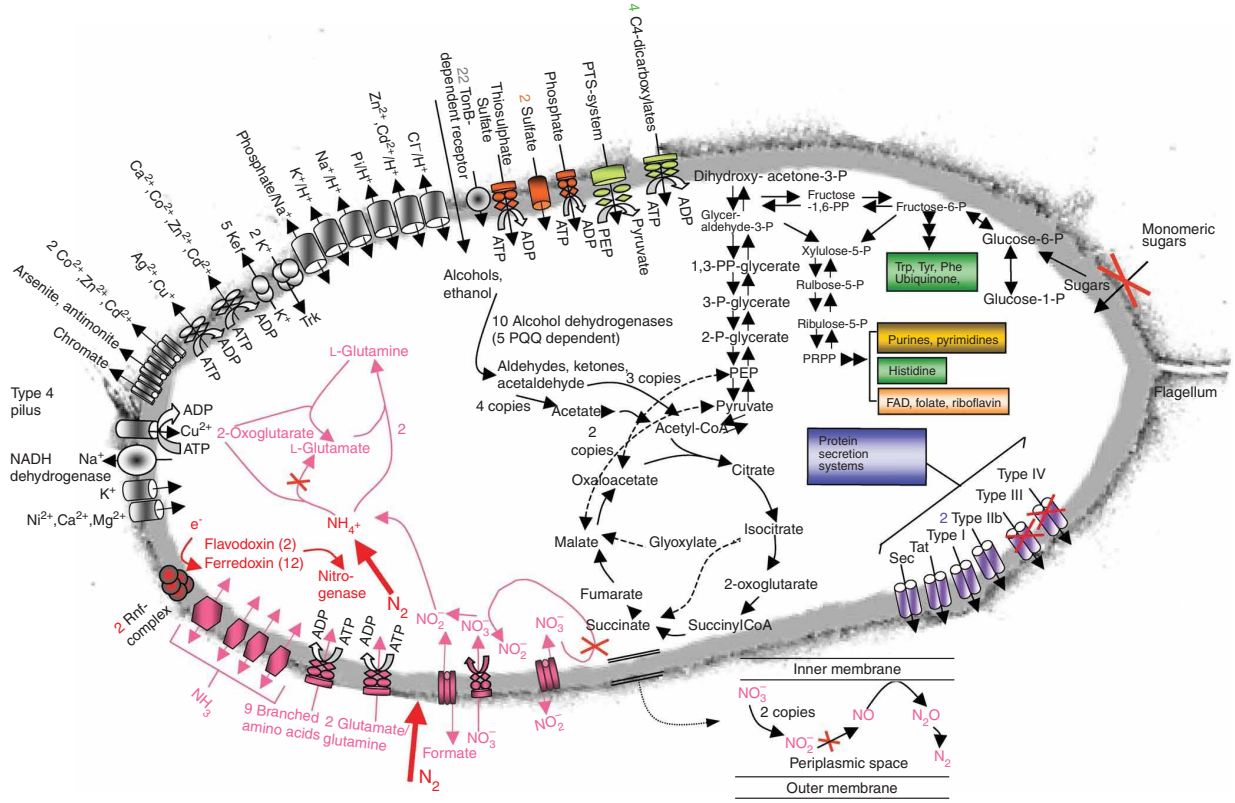


Figure 4 Overview of physiological features of *Azoarcus* sp. strain BH72. Depicted are carbohydrate metabolism, selected transporters and catabolic pathways for organic acids, nitrogen assimilatory and dissimilatory pathways, inorganic ion transport systems, and protein translocation systems. Crosses indicate pathways or reactions that are apparently not present in strain BH72. Details of metabolic features: the genome does not contain genes required for a functional Entner-Doudoroff pathway or the oxidative branch of the pentose phosphate pathway, however, the nonoxidative pentose phosphate pathway is complete. All enzymes required for glycolysis via the Embden-Meyerhoff pathway, gluconeogenesis and the TCA cycle are present, including genes encoding a typical phosphoenolpyruvate: sugar phosphotransferase system (PTS), but no specific outer membrane transporter for carbohydrates. Four complete and two incomplete copies of the TRAP-transport system (DctPQM) for C4-dicarboxylates are present. C4-dicarboxylates are probably metabolized via the glyoxylate shunt pathway, malate being decarboxylated by the malic enzyme (MaeB, two gene copies). Ten putative alcohol dehydrogenases are encoded in the genome (see also **Supplementary Table 1** online), five of which are PQQ-dependent enzymes not present in strain EbN1.

nitrite transport and reduction are present in strain BH72 in one copy (*nasFED*, *nirC*, *nasA*, *nasC*, *nirBD*). In contrast to strain EbN1, strain BH72 cannot conduct denitrification to N_2, as genes required for a nitrite reductase are missing in strain BH72. However, genes for subsequent denitrification reactions (*norCBQD* and *nosRZDYFLX* encoding reductases for nitric oxide and nitrous oxide, respectively) are present. The periplasmic localization of the nitrate reductase components (duplicated *nap* operon) and detoxification of NO_2^- via an NO_3^-/NO_2^- antiporter (NarK) might decrease toxicity of nitrite. In contrast, the denitrification pathway is complete in strain EbN1 for which denitrification appears to be a typical feature[10].

Iron-transport related proteins

In Gram-negative bacteria, TonB-dependent, outer-membrane proteins are responsible for the specific uptake of ferric-siderophore complexes, high-affinity iron chelators. They are also important for perception of environmental signals and are associated with pathogenicity of plant pathogens[19]. Strain BH72 possesses 22 genes encoding proteins related to iron transport (**Supplementary Table 1** online), twice as many as described for strain EbN1 and even more than other N_2-fixing endosymbionts (*Bradyrhizobium japonicum*, 13; *Sinorhizobium meliloti*, 9; and *Mesorhizobium loti*, 1). Two genes (*azo2156*, *azo3836*) are not even present in the *P. fluorescens* Pf5 genome, a plant-associated bacterium known for its capacity to produce and take up a wide range of siderophores[20], which contains 45 such genes. Although putative receptors for hydroxamate- and catechol-type siderophores, ferricitrate, vitamin B12, colicins and unknown substances are present, there was no evidence for biosynthetic pathways for known hydroxamate or catecholate siderophores[21]. Moreover, production of siderophores was not detected experimentally (**Supplementary Fig. 1a** online). Apparently, this strain is highly adapted to obtaining chelated iron from other sources, as fungi and monocotyledonous plants also produce siderophores[22]. The high number of putative receptors suggests a role not only for rhizosphere competence of strain BH72, but also for biocontrol.

Plant-associated lifestyle

Surface characteristics of bacteria are important factors for recognition by and interaction with the host. Several gene clusters putatively related to surface components of strain BH72 are lacking in strain EbN1, or are more highly related to genes of plant-associated or pathogenic bacteria.

Type IV pili are among the few factors known to affect endophytic colonization of grass diazotrophs[23]. Establishment of microcolonies on roots and fungal mycelium, and systemic spreading in rice are mediated by type IV pili[23]. The strain BH72 genome harbors 41 genes encoding proteins putatively involved in pilus assembly and regulation, whereas only 30 such genes were found in strain EbN1. Genes highly similar in both species encode proteins with conserved function such as assembly or regulation (for example, PilBCD-PilF-PilM-NOPQ-PilTU-PilZ or PilSR-PilGHIL). Other pilus proteins related mostly to phytopathogenic bacteria might be either pseudopilins involved in secretion (PilV/W/X) or putative tip adhesins (PilY1A/B, 31% and 39% sequence identity to *Ralstonia solanacearum* and *Xylella fastidosa*, respectively) that are lacking even in strain EbN1 and might be characteristic for interaction with plant surfaces (**Supplementary Table 1** online).

Other cell surface components that are often involved in recognition or virulence of pathogens are lipopolysaccharides, exopolysaccharides and capsular material. Intriguingly, many gene products putatively involved in their synthesis in strain BH72 are not highly related to those of the soil isolate EbN1, but to proteins of plant symbionts, pathogens or gamma- or alpha-proteobacteria (**Supplementary Table 1** online). Several genes of strain BH72 are similar to the *gum* operon for exopolysaccharide production in phytopathogenic *Xanthomonas campestris*[24]; genes encoding putative glycosyl transferases have considerable similarity to the rhizobial *pss* gene cluster (**Supplementary Table 1** online, 20–23), which is involved in exopolysaccharide polymerization, translocation and thus in plant-microbe interaction; a gene cluster related mainly to lipopolysaccharide synthesis is most similar to genes of gamma-proteobacteria including pathogens; these clusters did not show sufficient synteny to support the assumption of a very recent gene transfer (**Fig. 3**).

Motility. Flagella are pivotal for motility, adhesion, biofilm formation and colonization of the host. Strain BH72 is highly motile by means of a polar flagellum. At least 48 genes were identified that are generally required for biosynthesis and function of flagella and chemotaxis. They are located in three different noncontiguous clusters and are mostly related to genes of other beta-proteobacteria and a few pathogens of the gamma-subgroup (**Supplementary Table 1** online). There are three genes encoding flagellins (*fliC1, fliC2, fliC3*) and two encoding flagellar motor proteins, suggesting an important role for motility in the plant-associated lifestyle. In contrast, the nonmotile strain EbN1 does not possess a complete flagellar regulon[10].

Secretion and communication. Several genes encoding potential protein secretion systems were identified in the genome of strain BH72 (**Fig. 4**) for a sec-dependent pathway, a signal recognition particle (SRP)-mediated translocation and a twin arginine translocation (Tat) system, all of them targeting proteins through the inner membrane. Secretion of proteins through the entire cell envelope seems to be limited to only three varieties of pathways. Genes were identified that encode one type I secretion system and one autotransporter. Two gene clusters were detected that encode a type II secretion-related system (type IIb secretion system[25]), consisting only of GspDEFG.

Two other secretion systems are common to plant-associated bacteria, type III and IV secretion systems, which transport a wide variety of effector proteins into the extracellular medium or into the cytoplasm of eukaryotic host cells and affect interaction[26–29]. Intriguingly, neither system is present in strain BH72, probably preventing the export of toxic proteins to the host.

'Quorum sensing' is a common way of bacteria to communicate with each other or hosts by means of autoinducers that accumulate in the extracellular environment in a cell density-dependent manner[30]. Although there is evidence that autoinducer-dependent gene regulation occurs in *Azoarcus* sp. strain BH72 (Böhm, M. & Reinhold-Hurek, B., unpublished data), this strain appears to escape the usual communication systems. Widespread autoinducers of Gram-negative bacteria are N-acyl homoserine lactones (AHLs)[30]. There is no evidence for genes encoding an AHL-based quorum-sensing system in strain BH72; genes encoding the autoinducer synthetase (LuxI/LasI–type) or the responsible cytoplasmic autoinducer receptor (LuxR/LasR-type) are lacking. Furthermore, different bacterial sensor strains detecting presence of short- or long-chain AHLs did not yield a positive response toward strain BH72 (**Supplementary Fig. 1b** online). Also genes encoding the autoinducer-2 synthetase LuxS[30] are lacking. Gram-positive bacteria usually use peptides as autoinducers[30]. Genes expected for this system were not detected in strain BH72 either.

Virulence and interaction factors. The strain BH72 genome stands out by the lack of obvious genes involved in production of toxins. Moreover, common hydrolytic enzymes that macerate plant cell wall polymers and thus contribute to a phytopathogenic lifestyle and plant damage are rare: pectinase-encoding genes are absent; only a few genes encode putative glycosidases (*palZ*, *spr1*, *ndvC*, *eglA*), some of them with transmembrane helices. Detection of genes for membrane-bound enzymes is in agreement with the observation that strain BH72 does not secrete cellulases into the culture medium, but shows activities of a cell surface-bound endoglucanase (EglA) and exoglycanase that also hydrolyses xylosides[16], a major component of primary cell walls in grasses. A low production of macerating enzymes is likely to contribute to compatibility with the plant host, however these hydrolases might assist in endophytic colonization, as shown for the endoglucanase EglA[31].

There is no genomic evidence for the central process in the rhizobium-legume symbiosis, the induction of nodulation. Common *nodABC* genes required for the biosynthesis of the Nod-factor backbone are not present in strain BH72; only a few genes show some sequence similarity to other *nod* or *nol* genes (**Supplementary Table 1** online). However, like other grass-associated microbes such as *Azospirillum* a gene similar to *nodD* is present, in rhizobia encoding a central regulator for flavonoid-inducible gene expression.

Some other gene clusters in strain BH72 are also interesting targets for putative roles in plant-microbe interactions. One cluster shows similarity to genes that are localized in the *sci*-genomic island, affecting virulence of human pathogens. The genomic organization shows remarkable synteny with genes of *P. aeruginosa* (**Fig. 3**), arguing for a more recent gene transfer. Interestingly some homologs are also present in rhizobia. Other noticeable gene clusters code for mainly conserved hypothetical proteins or metabolism-related proteins that have orthologs in rhizobia, or of regulatory proteins (ColRS) important for root colonization of commensals (**Supplementary Table 1** online, no. 177–185, 188–196, 198–200).

DISCUSSION

The complete genome sequence of *Azoarcus* sp. strain BH72 offers insights into genomic strategies for an endophytic life style, and allows identification of various features that may contribute to their interaction with plants. The strain appears to be adapted to a relatively stable, low-stress microenvironment, since its genome contains remarkably few phage- or transposon-related genes in comparison to many soil bacteria or pathogens, indicating a low plasticity of the

genome. The lack of the typical communication system of Gram-negative bacteria based on AHLs also argues for a rather exclusive microhabitat.

Strain BH72 appears to be disarmed compared to plant pathogens by its in ability to metabolize carbohydrates coupled with the lack of a massive occurrence of cell-wall degrading enzymes. Moreover, this bacterium lacks known toxins and type III and IV secretion systems that are typically used by pathogens to transport effector molecules to their host. This might be instrumental for avoiding damage to the plant host despite a dense internal colonization, and only the small set of hydrolases identified may be required for penetrating into the plant tissue. Some specific features of nodule symbionts are also lacking, like most *nod*/*nol* genes required for nodule induction.

Genome comparison with the *Azoarcus*-related, nondiazotrophic, nonendophytic soil bacterium strain EbN1 revealed features likely to be important for plant-microbe interaction. Strain BH72 appears to be highly adapted to environments of low nitrogen availability, N_2-fixation playing a key role in its ecology. Several gene clusters that were lacking in strain EbN1 or were highly similar to genes of plant-associated or pathogenic bacteria were related to cell surface components that are often involved in recognition—gene products participating in the synthesis of exopolysaccharide, lipopolysaccharide, type IV pilus tips, the flagellar and chemotaxis apparatus. Further targets for studying interaction mechanisms were also identified by comparative genomics, such as virulence-related *sci* genes or genes encoding conserved hypothetical proteins shared with nodule symbionts. A large and diverse set of TonB-dependent receptors (22) might play a role in iron acquisition and biocontrol. In future functional genomic analyses, the role of these target genes for host compatibility will be elucidated, which is crucial for a wider agro-biotechnological application of N_2-fixing endophytes.

METHODS

Whole genome shotgun sequencing. DNA shotgun libraries with insert sizes of 1 kb and 2–3 kb in pGEM-T (Promega), and 8-kb fragments in pTrueBlue-rop (MoBiTec) vectors were constructed by MWG Biotech. Plasmid clones were end-sequenced on ABI 3700 sequencers (ABI) by MWG Biotech AG. Base-calling was carried out using PHRED[32,33]. High-quality reads were defined by a minimal length of 250 bp with an averaging quality value of ≥ 20. Finally, 60,715 high-quality reads, a total of 39,266 (5.26 ×), 18,070 (2.32 ×) and 3,379 (0.40 ×) end sequences (×'s indicate genome equivalents) from libraries with 1-kb, 2- to 3-kb and 8-kb inserts, respectively, were obtained.

Sequence assembly and assembly validation. Basecalling, quality control and elimination of vector DNA sequences of the shotgun-sequences were performed by using the software package BioMake (Bielefeld University) as previously described[34]. Sequence assembly was performed by using the PHRAP assembly tool (http://www.phrap.org). The CONSED/AUTOFINISH software package[35,36], supplemented by the in-house tool BACCardI[37], was used for the finishing of the genome sequence.

For gap closure, a BAC library with inserts of ~90 kb in pBeloBAC11 was constructed and BAC contigs were assembled[12]. Remaining gaps of the whole genome shotgun assembly were closed by sequencing on shotgun and BAC clones carried out by IIT GmbH on LI-COR 4200L and ABI 377 sequencing machines. So that we would obtain a high quality genome sequence, all regions of the consensus sequence were polished to at least phred40 quality by primer walking. Collectively, 1,374 sequencing reads were added to the shotgun assembly for finishing and polishing of the genomic sequence. Repetitive elements, that is, rRNA operons, were sequenced completely by primer walking on BAC clones. For assembly validation, a fosmid library with inserts of ~35–38 kb was constructed by IIT GmbH using the EpiFOS Fosmid Library Production Kit (Epicentre). End-sequencing of 672 fosmids was carried out on ABI 377 and MegaBACE 1000 (Amersham Biosciences) sequencing machines by IIT GmbH and on ABI 3730XL DNA analyzers by the sequencing group of the Max Planck Institut für Molekulare Genetik. For assembly validation, fosmid end sequences were mapped onto the genome sequence by employing the BACCardI tool.

Genome analysis and annotation. In a first step automatic gene prediction and annotation were performed using the genome annotation system GenDB 2.0 (ref. 38) as previously described[14]. In a second step the coding sequences (CDS) prediction was validated: a position weight matrix (PWM) was generated to score all translation starts, and visualization of CDS was done using GeneQuest (DNASTAR Inc.). Reinspection of starts was coupled to recomputing of homology (BlastP) and assessment of function. In this way 4.3% more ORFs were detected, and 15.5% of the start sites were changed in comparison to the prediction obtained by a combined GLIMMER and CRITICA approach[39]. Intergenic regions were checked again for CDS missed probably by the automatic annotation using the BLAST programs[40]. During manual annotation, the following criteria were applied: (i) for hypothetical proteins, amino acid identities to other proteins were less than 30% over the entire length of the protein; (ii) for conserved hypothetical proteins, amino acid identities to proteins of unknown function were more than 30%; (iii) for proteins with putative or probable functions, sequence identities to named gene products were >20% or 40%, respectively.

Genomic comparison. For comparative analyses, the annotated genome sequence of *Azoarcus* strain EbN1 (acc. nos. CR555306, CR555307, CR555308) was imported into GenDB. Comparisons of chromosomal sequences were carried out with GenDB[38].

Detection of regions with atypical G+C content. For detection of anomalous gene clusters or putative pathogenicity islands in bacterial genomes, the Pai-Ida program 1.1 (http://compbio.sibsnet.org/projects/pai-ida/) based on an iterative discriminant analysis[41] was used.

Database submission. The nucleotide sequence of the *Azoarcus* sp. strain BH72 chromosome was submitted to EBI under accession number AM406670–*Azoarcus* sp. BH72.

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

ACKNOWLEDGMENTS

The authors grateful thank all the people involved in this project. This work was supported by grants of the German Federal Ministry of Education and Research (BMBF) (031U113D, 031U213D and 0313105) in the frame of the GenoMik network "Genome Research on Bacteria Relevant for Agriculture, Environment and Biotechnology" and the Fonds der Chemischen Industrie (to R.T.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

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

1. Barac, T. *et al.* Engineered endophytic bacteria improve phytoremediation of water-soluble, volatile, organic pollutants. *Nat. Biotechnol.* **22**, 583–588 (2004).
2. Hurek, T., Handley, L., Reinhold-Hurek, B. & Piché, Y. *Azoarcus* grass endophytes contribute fixed nitrogen to the plant in an unculturable state. *Mol. Plant Microbe Interact.* **15**, 233–242 (2002).
3. Sevilla, M., Burris, R.H., Gunapala, N. & Kennedy, C. Comparison of benefit to sugarcane plant growth and 15N2 incorporation following inoculation of sterile plants with Acetobacter diazotrophicus wild-type and *nif*- mutant strains. *Mol. Plant-Microbe Interact.* **14**, 358–366 (2001).
4. Hurek, T. & Reinhold-Hurek, B. *Azoarcus* sp. strain BH72 as a model for nitrogen-fixing grass endophytes. *J. Biotechnol.* **106**, 169–178 (2003).
5. Reinhold, B., Hurek, T., Niemann, E-G. & Fendrik, I. Close association of *Azospirillum* and diazotrophic rods with different root zones of Kallar grass. *Appl. Environ. Microbiol.* **52**, 520–526 (1986).
6. Hurek, T., Reinhold-Hurek, B., Van Montagu, M. & Kellenberger, E. Root colonization and systemic spreading of *Azoarcus* sp. strain BH72 in grasses. *J. Bacteriol.* **176**, 1913–1923 (1994).
7. Miché, L., Battistoni, F., Gemmer, S. & Belghazi, M. Reinhold-Hurek, B. Differential colonization and up-regulation of jasmonate-inducible defence proteins in roots of *Oryza sativa* cultivars upon interaction with the endophyte *Azoarcus* sp. *Mol. Plant-Microbe Interact.* **19**, 502–511 (2006).

8. Egener, T., Hurek, T. & Reinhold-Hurek, B. Endophytic expression of *nif* genes of *Azoarcus* sp. strain BH72 in rice roots. *Mol. Plant-Microbe Interact.* **12**, 813–819 (1999).
9. Reinhold-Hurek, B., Tan, Z. & Hurek, T. in Bergey's Manual of Systematic Bacteriology. Vol. 2. (ed. G.M. Garrity) 890–901, (Springer Verlag, New York, 2005).
10. Rabus, R. *et al.* The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. *Arch. Microbiol.* **183**, 27–36 (2005).
11. Reinhold-Hurek, B. & Hurek, T. Life in grasses: diazotrophic endophytes. *Trends Microbiol.* **6**, 139–144 (1998).
12. Battistoni, F. *et al.* Physical map of the *Azoarcus* sp. strain BH72 genome based on a bacterial artificial chromosome (BAC) library as a platform for genome sequencing and functional analysis. *FEMS Microbiol. Lett.* **249**, 233–240 (2005).
13. James, E.K. & Olivares, F.L. Infection and colonization of sugar cane and other graminaceous plants by endophytic diazotrophs. *Crit. Rev. Plant Sci.* **17**, 77–119 (1998).
14. Thieme, F. *et al.* Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris pv. vesicatoria* revealed by the complete genome sequence. *J. Bacteriol.* **187**, 7254–7266 (2005).
15. Reinhold-Hurek, B. *et al. Azoarcus* gen. nov., nitrogen-fixing proteobacteria associated with roots of Kallar grass (*Leptochloa fusca* (L.) Kunth) and description of two species *Azoarcus indigens* sp. nov. and *Azoarcus communis* sp. nov. *Int. J. Syst. Bacteriol.* **43**, 574–584 (1993).
16. Reinhold-Hurek, B., Hurek, T., Claeyssens, M. & Van Montagu, M. Cloning, expression in *Escherichia coli*, and characterization of cellulolytic enzymes of *Azoarcus* sp., a root-invading diazotroph. *J. Bacteriol.* **175**, 7056–7065 (1993).
17. Yurgel, S.N. & Kahn, M.L. Dicarboxylate transport by rhizobia. *FEMS Microbiol. Rev.* **28**, 489–501 (2004).
18. Hurek, T., Reinhold, B., Fendrik, I. & Niemann, E.G. Root-zone-specific oxygen tolerance of *Azospirillum* spp. and diazotrophic rods closely associated with Kallar grass. *Appl. Environ. Microbiol.* **53**, 163–169 (1987).
19. Koebnik, R. TonB-dependent trans-envelope signalling: the exception or the rule? *Trends Microbiol* **13**, 343–347 (2005).
20. Paulsen, I.T. *et al.* Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat. Biotechnol.* **23**, 873–878 (2005).
21. Crosa, J.H. & Walsh, C.T. Genetics and assembly line enzymology of siderophore biosynthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **66**, 223–249 (2002).
22. Crowley, D.E., Wang, Y.C., Reid, C.P.P. & Szaniszlo, P.J. Mechanism of iron acquisition from siderophores by microorganism and plants. *Plant Soil* **130**, 179–198 (1991).
23. Dörr, J., Hurek, T. & Reinhold-Hurek, B. Type IV pili are involved in plant-microbe and fungus-microbe interactions. *Mol. Microbiol.* **30**, 7–17 (1998).
24. Katzen, F. *et al. Xanthomonas campestris* pv. campestris *gum* mutants: effects on xanthan biosynthesis and plant virulence. *J. Bacteriol.* **180**, 1607–1617 (1998).
25. Filloux, A. The underlying mechanisms of type II protein secretion. *Biochim. Biophys. Acta* **1694**, 163–179 (2004).
26. He, S.Y., Nomura, K. & Whittam, T.S. Type III protein secretion mechanism in mammalian and plant pathogens. *Biochim. Biophys. Acta* **1694**, 181–206 (2004).
27. Krause, A., Doerfel, A. & Göttfert, M. Mutational and transcriptional analysis of the type III secretion system of *Bradyrhizobium japonicum*. *Mol. Plant-Microbe Interact.* **15**, 1228–1235 (2002).
28. Christie, P.J. Type IV secretion: the *Agrobacterium* VirB/D4 and related conjugation systems. *Biochim. Biophys. Acta* **1694**, 219–234 (2004).
29. Hubber, A., Vergunst, A.C., Sullivan, J.T., Hooykaas, P.J. & Ronson, C.W. Symbiotic phenotypes and translocated effector proteins of the *Mesorhizobium loti* strain R7A VirB/D4 type IV secretion system. *Mol. Microbiol.* **54**, 561–574 (2004).
30. Camilli, A. & Bassler, B.L. Bacterial small-molecule signaling pathways. *Science* **311**, 1113–1116 (2006).
31. Reinhold-Hurek, B., Maes, T., Gemmer, S., Van Montagu, M. & Hurek, T. An endoglucanase is involved in infection of rice roots by the not cellulose-metabolizing endophyte *Azoarcus* sp. BH72. *Mol. Plant-Microbe Interact.* **19**, 181–188 (2006).
32. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
33. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
34. Kaiser, O. *et al.* Whole genome shotgun sequencing guided by bioinformatics pipelines–an optimized approach for an established technique. *J. Biotechnol.* **106**, 121–133 (2003).
35. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
36. Gordon, D., Desmarais, C. & Green, P. Automated finishing with autofinish. *Genome Res.* **11**, 614–625 (2001).
37. Bartels, D. *et al.* BACCardI–a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. *Bioinformatics* **21**, 853–859 (2005).
38. Meyer, F. *et al.* GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31**, 2187–2195 (2003).
39. McHardy, A.C., Goesmann, A., Pühler, A. & Meyer, F. Development of joint application strategies for two microbial gene finders. *Bioinformatics* **20**, 1622–1631 (2004).
40. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
41. Tu, Q. & Ding, D. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiol. Lett.* **221**, 269–275 (2003).
42. Reinhold-Hurek, B. & Shub, D.A. Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature* **357**, 173–176 (1992).

Paper III

Published online October 7, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 17 ***5691–5702***
doi:10.1093/nar/gki866

The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes

Ross Overbeek[1], Tadhg Begley[16], Ralph M. Butler[10], Jomuna V. Choudhuri[3], Han-Yu Chuang[17], Matthew Cohoon[12], Valérie de Crécy-Lagard[13], Naryttza Diaz[3], Terry Disz[12,], Robert Edwards[1,7,8], Michael Fonstein[1,18], Ed D. Frank[2], Svetlana Gerdes[1], Elizabeth M. Glass[2], Alexander Goesmann[3], Andrew Hanson[14], Dirk Iwata-Reuyl[15], Roy Jensen[5], Neema Jamshidi[17], Lutz Krause[3], Michael Kubal[12,], Niels Larsen[11], Burkhard Linke[3], Alice C. McHardy[3], Folker Meyer[3], Heiko Neuweger[3], Gary Olsen[9], Robert Olson[12], Andrei Osterman[1,8], Vasiliy Portnoy[17], Gordon D. Pusch[1], Dmitry A. Rodionov[6], Christian Rückert[4], Jason Steiner[17], Rick Stevens[2,12], Ines Thiele[17], Olga Vassieva[1], Yuzhen Ye[8], Olga Zagnitko[1] and Veronika Vonstein[1,*]

[1]Fellowship for Interpretation of Genomes, 15W155 81st Street, Burr Ridge, IL 60527, USA, [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA, [3]Center for Biotechnology, [4]International NRW Graduate School in Bioinformatics & Genome Research, Institute for Genome Research, Bielefeld University, 33594 Bielefeld, Germany, USA, [5]Emerson Hall, University of Florida, PO Box 14425, Gainesville, FL 32604, USA, [6]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, [7]Center for Microbial Sciences, San Diego State University, San Diego, CA 92813, USA, [8]The Burnham Institute, San Diego CA 92037, USA, [9]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, [10]Computer Science Dept, Middle Tennessee State University, Murfreesboro, TN 37132, USA, [11]Danish Genome Institute, Gustav Wieds vej 10 C, DK-8000 Aarhus C, Denmark, [12]Computation Institute, University of Chicago, Chicago, IL 60637, USA, [13]Departments of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611, USA, [14]Department of Horticultural Science, University of Florida, Gainesville, FL 32611, USA, [15]Department of Chemistry, Portland State University, Portland, OR 97207, USA, [16]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY14853, USA, [17]University of California, San Diego, CA 92093, USA and [18]Cleveland BioLabs, Inc., Cleveland, OH 44106, USA

Received June 9, 2005; Revised and Accepted September 8, 2005

ABSTRACT

The release of the 1000th complete microbial genome will occur in the next two to three years. In anticipation of this milestone, the Fellowship for Interpretation of Genomes (FIG) launched the Project to Annotate 1000 Genomes. The project is built around the principle that the key to improved accuracy in high-throughput annotation technology is to have experts annotate single subsystems over the complete collection of genomes, rather than having an annotation expert attempt to annotate all of the genes in a single genome. Using the subsystems approach, all of the genes implementing the subsystem are analyzed by an expert in that subsystem. An annotation environment was created where populated subsystems are curated and projected to new genomes. A portable notion of a *populated subsystem* was defined, and tools developed for exchanging and curating these objects. Tools were also developed to resolve conflicts between populated subsystems. The SEED is the first annotation environment that supports this model of annotation. Here, we describe the subsystem approach, and offer the first release of our growing library of populated subsystems. The initial release of data includes 180 177 distinct proteins

*To whom correspondence should be addressed. Tel: +1 630 325 4178; Fax: +1 630 325 4179; Email: Veronika@theFIG.info

with 2133 distinct functional roles. This data comes from 173 subsystems and 383 different organisms.

INTRODUCTION

In the 10 years since the first complete bacterial genome was released in 1995 (1) there has been an exponential growth in the number of complete genomes sequenced. More than 200 complete genomes have been released, and based on past growth we anticipate that the 1000th genome will be sequenced at some point during 2007 (Figure 1). This rapid release of data reinforces the need for high-throughput annotation systems that provide reliable and accurate results.

In response to these challenges the Fellowship for Interpretation of Genomes (FIG) launched the 'Project to Annotate a 1000 Genomes'. The Project embodies a specific strategic view of how to approach high-throughput annotation: the effort is organized around subsytem experts, individuals who master the details of a specific subsystem and then analyze and annotate the genes that make up that given subsystem over an entire collection of genomes.

We argue that a subsystems based approach provides many benefits compared to more traditional techniques of genome annotation:

(i) The analysis of a single subsystem over a large collection of genomes produces far more accurate annotations than the common approach of annotating the genes within a single organism. In fact the usual 'gene-by-gene' approach ensures that in most cases the individual annotating an entire genome lacks specific expertise related to the role of each gene.

(ii) The annotation of protein families rather than an organism at a time brings to bear specialized expertise and consequently leads to improvements over 'gene-by-gene' annotations of one genome. Just as the analysis of families offers a significant improvement over the annotation of individual genes, the analysis of *sets of related protein families* (i.e. those containing genes that make up a single biological subsystem) is more productive than the analysis of single families in isolation. Indeed, the fact that 'The presence or absence of metabolic pathways and structures provides a context that makes protein annotation far more reliable' (2) has now become clearly established.

(iii) It is both more straightforward and less error prone to automatically project annotations from a set of populated subsystems covering a diverse set of organisms than to project individual annotations using the existing automated pipelines. This is leading to the development of rule-based extension systems that will quite probably achieve superior accuracy (http://www.ebi.ac.uk/swissprot/Publications/dagstuhl.html).

(iv) A collection of annotations organized around specific subsystems covering a large number of diverse organisms represents a central resource for other bioinformatics efforts such as metabolic reconstruction, stoichiometric modeling and gene discovery (3).

This paper describes the subsystem-based approach to high-throughput genome annotation. The broad concepts of this approach are described and several examples of annotated subsystems are provided. Supplementary online material consisting of 173 subsystems has been released. Additionally, our open-source software for their creation and curation is provided.

WHAT IS A SUBSYSTEM

A *subsystem* is a set of *functional roles* that together implement a specific biological process or structural complex

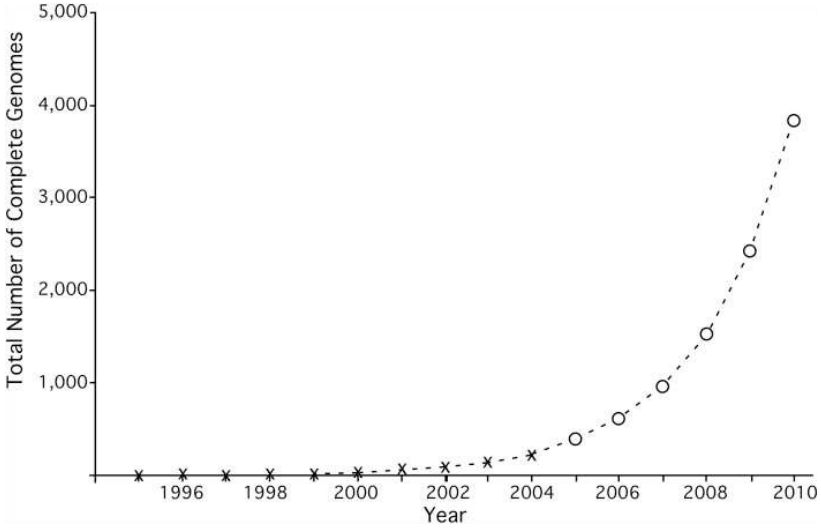


Figure 1. Accumulation of complete archaeal and bacterial genome sequences at NCBI 1994–2004, and prediction of the release of genomes through 2010. Data from http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi was extracted and plotted by year as shown with the crosses. Data from 2004–2010 is projected by the power law and is represented by open circles. At the current rate of growth, the 1000th complete microbial genome will be released in late 2007 or early 2008.

Table 1. Glossary

Annotation	An unstructured text string associated with specific genes and/or proteins.
Clearinghouse	A site for publish-request type peer-to-peer exchange of subsystems in a system independent manner.
Functional role	An abstract function that a protein performs. Subsystems developers specify a single, precise text string to represent each functional role.
Functional variants	Different combinations of functional roles that represent distinct operational forms of a subsystem.
Missing gene	A gene, that is predicted to be present in the genome of an organism but has not been identified yet.
Populated subsystem	A subsystem along with a spreadsheet in which each column represents a functional role for the subsystem, each row represents a specific genome, and each cell contains those genes from the specific organism that have a subsystem connection to the specific functional role.
Product name	A short text string used to represent the function of the protein encoded by a gene. No constraints are placed on the strings used as product names, and it is common to see the same abstract function denoted by numerous similar expressions.
Protein family	A collection of proteins that were grouped by a curator. Proteins may be grouped based on domain structure, similarity, or some other characteristic. Proteins within a family may implement the same or multiple functional roles.
Subsystem	A Subsystem is a collection of functional roles, which together implement a specific biological process or structural complex. There is no distinction between metabolic subsystems and non-metabolic subsystems.
Subsystem connections	The set of functional roles that tie protein-encoding genes to different subsystems. Most protein encoding genes currently have a single subsystem connection.
Variant code	Numeric codes used to distinguish different functional variants.

(Table 1). A subsystem may be thought of as generalization of the term *pathway*. Thus, just as glycolysis is composed of a set of functional roles (glucokinase, glucose-6-phosphate isomerase and phosphofuctokinase, etc.) a complex like the ribosome or a transport system can be viewed as a collection of functional roles. In practice, we put no restriction on how curators select the set of functional roles they wish to group into a subsystem, and we find subsystems being created to represent the set of functional roles that make up pathogenicity islands, prophages, transport cassettes and complexes (although many of the existing subsystems do correspond to metabolic pathways). The concept of *populated subsystem* is an extension of the basic notion of subsystem—it amounts to a subsystem along with a spreadsheet depicting the exact genes that implement the functional roles of the subsystem in specific genomes. The populated subsystem specifies which organisms include operational variants of the subsystem and which genes in those organisms implement the functional roles that make up the subsystem. Each column in the spreadsheet corresponds to a functional role from the subsystem, each row represents a genome, and each cell identifies the genes within the genome that encode proteins which implement the specific functional role within the designated genome (Figure 2).

The act of populating the subsystem amounts to adding rows (i.e. genomes) to the spreadsheet.

Since these concepts are fundamental to our discussion we are illustrating them in Figure 2.

Note that each row in the spreadsheet has an associated variant code. The set of roles that make up the example subsystem include all of the functional roles needed to encode three common variants of the pathway. The variant codes distinguish three alternative means of converting N-formimino-L-glutamate to L-glutamate.

We have adhered to the position that experts encoding subsystems must decide exactly which functional roles to include (and exactly how to express each functional role), as well as what variant codes to use. We have restricted the use of two variant codes: 0 to represent *work in progress* and −1 to represent *no operational variant*.

A FRAMEWORK FOR DEVELOPING A PRECISE VOCABULARY FOR FUNCTIONAL ROLES

Controlled vocabularies have often been proposed in computer-assisted annotations and data mining (4,5). Subsystems technology supports the definition of a controlled vocabulary for gene function. Domain experts, by defining the functional roles that make up the subsystems that they curate, impose a precise vocabulary for assignment of function to the genes that implement the subsystem. Since the term 'gene function' has come to have several meanings, it is important to distinguish between four concepts:

(i) A *functional role* is an abstract function such as 'Aspartokinase (EC 2.7.2.4)'. Subsystems are sets of such abstract functions.

(ii) The notion of *product name* refers to a short text string that someone has used to represent the function of the protein encoded by a gene. There are no constraints on the strings used as product names, and it is common to see the same abstract function denoted by numerous similar expressions such as 'Aspartokinase, Aspartokinase II, aspartate kinase' etc.

(iii) By the term *protein family* we mean a collection of proteins that have been grouped by some curation team. The UniProt effort is producing one particularly valuable collection of families. Within that effort, the protein family represents a set of proteins that share a common domain structure. That is, they may actually implement the same or multiple functional roles. Within our work, there is no explicit concept of protein family; the closest notion would be 'the set of genes within a single column of the spreadsheet in a populated subsystem'. However, a single column often contains proteins with distinct domain structure (e.g. both unifunctional and multifunctional proteins often occur within a single column), and in some cases genes encoding non-homologous proteins, which implement a single function have been included within a single column. We have developed tools to support comparison between protein families from a variety of sources and the proteins encoded by the genes in a

POPULATED*SUBSYSTEM

A

Subsystem: Histidine Degradation		
1	**HutH**	Histidine ammonia-lyase (EC 4.3.1.3)
2	**HutU**	Urocanate hydratase (EC 4.2.1.49)
3	**HutI**	Imidazolonepropionase (EC 3.5.2.7)
4	**GluF**	Glutamate formiminotransferase (EC 2.1.2.5)
5	**HutG**	Formiminoglutamase (EC 3.5.3.8)
6	**NfoD**	N-formylglutamate deformylase (EC 3.5.1.68)
7	**ForI**	Formiminoglutamic iminohydrolase (EC 3.5.3.13)

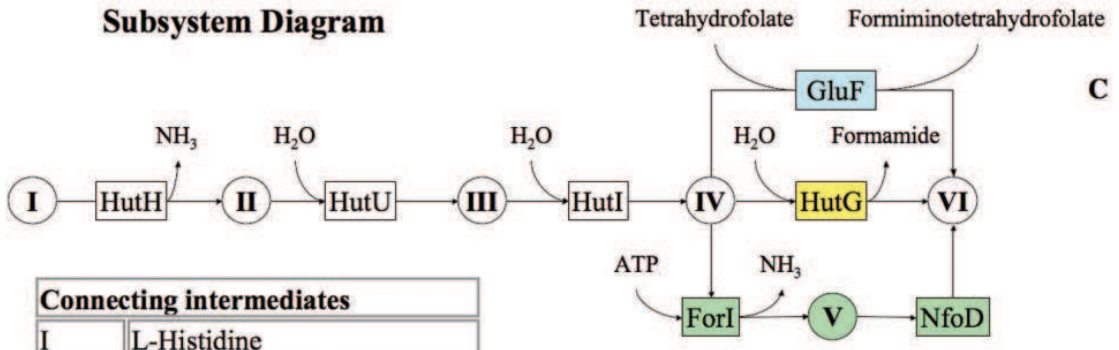
B

Subsystem Spreadsheet

Organism	Variant	HutH	HutU	HutI	GluF	HutG	NfoD	ForI
Caulobacter crescentus	1	P58082	Q9A9M1	P58079			Q9A9M0	Q9A9L9
Pseudomonas putida	1	Q88CZ7	Q88CZ6	Q88CZ9			Q88D00	Q88CZ3
Xanthomonas campestris	1	Q8PAA7	P58988	Q8PAA6			Q8PAA8	Q8PAA5
Halobacterium sp.	2	Q9HQD5	Q9HQD8	Q9HQD6		Q9HQD7		
Deinococcus radiodurans	2	Q9RZ06	Q9RZ02	Q9RZ05		Q9RZ04		
Bacillus subtilis	2	P10944	P25503	P42084		P42068		
Bacteroides thetaiotaomicron	3	Q8A4B3	Q8A4A9	Q8A4B1	Q8A4B0			
Desulfotela psychrophila	3	gi51246205	gi51246204	gi51246203	gi51246202			

C

Subsystem Diagram



Connecting intermediates	
I	L-Histidine
II	Urocanate
III	4-Imidazolone-5-Propanoate
IV	N-Formimino-L-Glutamate
V	N-Formyl-L-Glutamate
VI	L-Glutamate

Figure 2. Subsystem and Populated Subsytem. The Histidine Degradation Subsystem was used as an example to demonstrate relevant terms. (**A**) The subsystem comprises of 7 functional roles (e.g. Histidine ammonia-lyase (EC 4.3.1.3), Urocanate hydratase (EC 4.2.1.49) etc.). Together with the spreadsheet it becomes the 'Populated subsystem'. (**B**) The Subsystem Spreadsheet is populated with genes from 8 organisms (simplified from the original subsystem) where each row represents one organism and each column one of the functional roles of the subsystem. Genes performing the specific functional role in the respective organism populate the respective cell. Gray shading of cells indicates proximity of the respective genes on the chromosomes. (**C**) The Subsystem Diagram illustrates the populated subsystem: key intermediates (circles with roman numerals), connected by enzymes (boxes with abbreviations matching the spreadsheet abbreviations) and reactions (arrows). There are three distinct variants of Histidine Degradation presented in this populated subsystem. Variant 1 (green shading) is present in *Caulobacter crescentus*, *Pseudomonas putida* and *Xanthomonas campestris*. N-Formimino-L-Glutamate (IV) is converted to L-Glutamate (VI) via N-Formyl-L-Glutamate (V) by enzymatic activities of Formiminoglutamic iminohydrolase (EC 3.5.3.13) (ForI) and of N-formylglutamate deformylase (EC 3.5.1.68) (NfoD). Variant 2 (yellow shading) is present in *Halobacterium sp.*, *Deinococcus radiodurans* and *Bacillus subtilis*. In this variant the conversion from intermediate IV to VI is performed by Formiminoglutamase (EC 3.5.3.8) (HutG). Variant 3 (blue shading) is present in *Bacteroides thetaiotaomicron* and *Desulfotela psychrophila*. Here the Glutamate formiminotransferase (EC 2.1.2.5) (GluF) performs the conversion from intermediate IV to VI.

single column in a populated subsystem. These comparisons are valuable but it is important to realize that we are producing sets of genes that encode proteins capable of implementing a single functional role, while the underlying restrictions on what make up a protein family often differ markedly from this notion.

(iv) The notation of *annotation* is often used to refer to an unstructured text string associated with specific genes and/or proteins.

To illustrate our use of these terms, consider the product name 'Lysine-sensitive aspartokinase III'. It implements the functional role 'Aspartokinase (EC 2.7.2.4)', which a curator has included in the subsystem 'Lysine_Biosynthesis_DAP_Pathway'. The curator may have well attached the annotation 'Cassan *et al*., 1986 Nucleotide sequence of lysC gene encoding the lysine-sensitive aspartokinase III of *Escherichia coli* K12. Evolutionary pathway leading to three isofunctional enzymes, *J. Biol. Chem*., 261, 1052–1057' for the respective E.coli K12 gene, justifying the use of this specific product name.

To this mix of concepts we add the notion *subsytem connection*. A gene can be connected to one or more functional roles, which induces connections to specific subsystems (those that contain the specific functional roles). In the example above it would be the connection to the subsystem 'Lysine_Biosynthesis_DAP_Pathway'.

Although product names often include special properties (e.g. 'thermostable' or 'lysine-sensitive'), and occasionally clues of function (e.g. '*similar to death associated protein kinase*'), subsystem connections unambiguously reference specific functional roles included in the definition of a subsystem.

Initially, the number of populated subsystems grew rapidly including numerous metabolic pathways, as well as non-metabolic subsystems ranging from flagella (http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=Flagellum&request=show_ssa, pathogenicity islands, http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=Mannose-sensitive_hemagglutinin_type_4_pilus&request=show_ssa), and secretory systems [http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=General_secretory_pathway_(Sec-SRP)_complex_(TC_3.A.5.1.1)&request=show_ssa] through complexes like the ribosome and proteosome. As both subsystems and the consequent subsystem connections matured there was considerable overlap between subsystems. Users developing subsystems on their own machines and sharing them through the clearinghouse exacerbated the differences in style, and hence conflicts between subsystems. For example, functional roles corresponding to the notion of *aconitase* exist in at least three distinct subsystems: the TCA cycle (http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=TCA_Cycle&request=show_ssa), the methylcitate cycle (http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=Methylcitrate_cycle&request=show_ssa), and glyoxylate synthesis (http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=Glyoxylate_Synthesis&request=show_ssa) developed independently by different curators. In at least one instance a curator wished to carefully distinguish three distinct forms of the enzyme. Initially each curator annotated the same protein-encoding genes with different functional roles, however this quickly became untenable—i.e. conflicts arose. To support uniform terminology required that the conflicts be detected, and be resolved by renaming functional roles to a consistent vocabulary employed consistently by all three subsystems. Rather than impose a centralized mechanism for resolving such conflicts, a completely decentralized approach was used.

To facilitate coordination and communication between end users, to aid with conflict resolution, and to eliminate redundancy, a multi-author website was developed using Wiki technology (http://www-unix.mcs.anl.gov/SEEDWiki/moin.cgi/MoinMoin). The subsystem bulletin board (http://www.theseed.org/wiki/moin.cgi/SubsystemBulletinBoard) provides an overview of the subsystems and highlights individual researcher's efforts. For a more detailed discussion of each of the subsystems, a Forum was developed using vBulletin technology (http://www.vbulletin.com/). The Forum (http://www.subsys.info) has subsystems separated by class, and each subsystem has a discussion arena for the deposition of comments, questions, suggestions and ideas. In addition to these resources, interactive conflict detection and resolution software was developed for the installation of subsystems in the SEED database.

Ultimately the success of our approach has been based on the good will and common desire to produce a consistent, precise vocabulary for functional roles, and we feel that this has worked well. It has produced a situation in which, at any given time, conflicts may exist because new subsystems are being developed or existing ones extended. But the attention of curators is being alerted to those instances by the development of tools that point to the conflicts. No centralized authority is being employed (although, in fact, on occasion curators do settle disagreements by consulting with outside experts). Conflicts can be of various types ranging from simple differences in spelling of functional roles to disagreements relating to specificity and numerous other issues. In all cases curators have reached settlements through discussions that lead to either consensus names or extended names. Once agreement has been reached and consistency established, changing the precise string of text that describes a functional role at some later point in time is trivial.

The result has been a vocabulary for functional roles that is precise, reasonably consistent, and rapidly improving. Our strategy for coupling this vocabulary with widely practiced ontologies such as GO will be to attach GO terms to each of the functional roles (inducing connections to genes via subsystem connections).

SUBSYSTEMS: A TECHNOLOGY INDEPENDENT OF ANNOTATION SYSTEMS

The subsystems technology described herein was developed with two primary goals in mind.

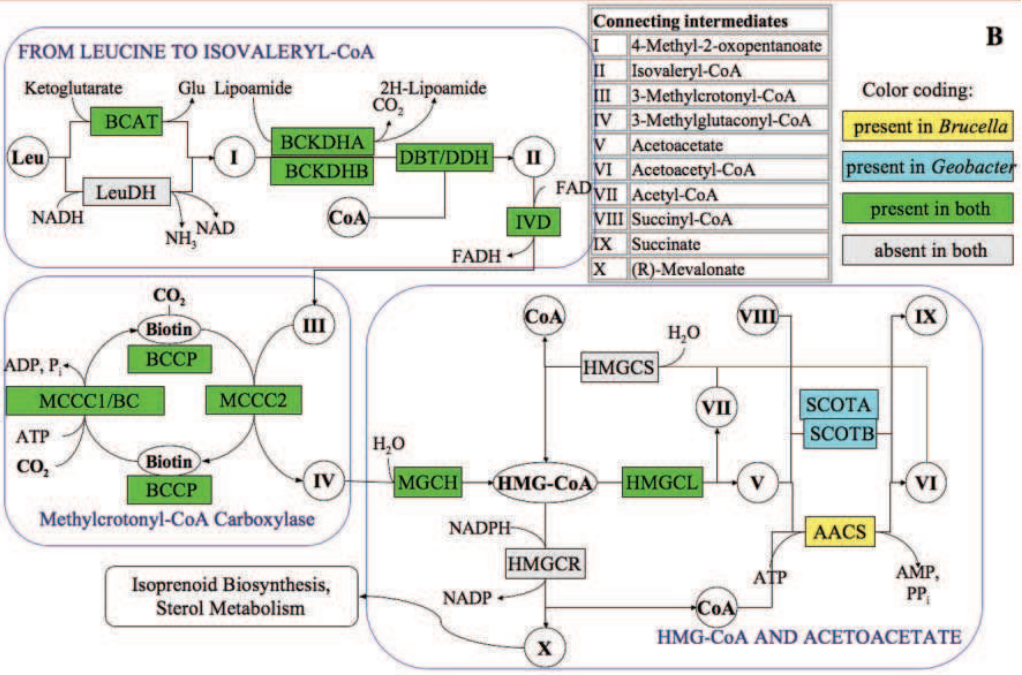
The first goal was to define a simple, portable text representation of a populated subsystem. This allowed populated subsystems to be exchanged, archived and updated over the Internet.

And the second goal to develop a *clearinghouse* where curators can publish populated subsystems for exchange with other users. The clearinghouse is available for direct querying from within a program (http://clearinghouse.theseed.org/) or via a web-browser (http://clearinghouse.theseed.org/clearinghouse_browser.cgi).

A

Subsystem: Leucine Degradation and HMG-CoA Metabolism		
1	**BCAT**	Branched-chain amino acid aminotransferase (EC 2.6.1.42)
2	**LDH**	Leucine dehydrogenase (EC 1.4.1.9)
3	**BCKDHA**	2-oxoisovalerate dehydrogenase alpha subunit (EC 1.2.4.4)
4	**BCKDHB**	2-oxoisovalerate dehydrogenase beta subunit (EC 1.2.4.4)
5	**DBT**	Lipoamide acyltransferase component of branched-chain alpha-keto acid dehydrogenase complex (EC
6	**DDH**	Dihydrolipoamide dehydrogenase (EC 1.8.1.4)
7	**IVD**	Isovaleryl-CoA dehydrogenase (EC 1.3.99.10)
8	**MCCC1**	Methylcrotonyl-CoA carboxylase biotin-containing subunit (EC 6.4.1.4)
9	**BC**	Biotin carboxylase of methylcrotonyl-CoA carboxylase (EC 6.3.4.14)
10	**BCCP**	Biotin carboxyl carrier protein of methylcrotonyl-CoA carboxylase
11	**MCCC2**	Methylcrotonyl-CoA carboxylase carboxyl transferase subunit (EC 6.4.1.4)
12	**MGCH**	Methylglutaconyl-CoA hydratase (EC 4.2.1.18)
13	**HMGCL**	Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4)
14	**AACS**	Acetoacetyl-CoA synthetase (EC 6.2.1.16)
15	**SCOTA**	Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit A (EC 2.8.3.5)
16	**SCOTB**	Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit B (EC 2.8.3.5)
17	**HMGCS**	Hydroxymethylglutaryl-CoA synthase (EC 2.3.3.10)
18	**HMGCR**	Hydroxymethylglutaryl-CoA reductase (EC 1.1.1.34)

B



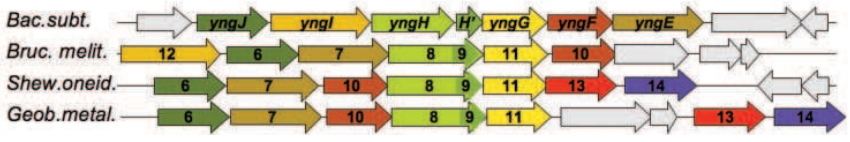
Connecting intermediates	
I	4-Methyl-2-oxopentanoate
II	Isovaleryl-CoA
III	3-Methylcrotonyl-CoA
IV	3-Methylglutaconyl-CoA
V	Acetoacetate
VI	Acetoacetyl-CoA
VII	Acetyl-CoA
VIII	Succinyl-CoA
IX	Succinate
X	(R)-Mevalonate

Color coding:
present in *Brucella*
present in *Geobacter*
present in both
absent in both

C

Genomes	Funct. variant	From Leucine to Isovaleryl-CoA						To Acetoacetate via HMG-CoA							Acetoacetate			HMG-CoA	
		BCAT	LeuDH	BCKDHA	BCKDHB	DBT	DDH	IVD (6)	MCCC1 (8/9)	BC (8)	BCCP (9)	MCCC2 (7)	MGCH (10)	HMGCL (11)	AACS (12)	SCOTA (13)	SCOTB (14)	HMGCS	HMGCR
Bac.subtil.	1	*ywaA*	*bcd*	*bkdA1*	*bkdA2*	*bkdB*	*yqiV*	*yngJ*		*yngH*	*yngH'*	*yngE*	*yngF*	*yngG*	*yngI*	*yxjD*	*yxjE*	*pksG*	
Bruc.melit.	2	+		+	+	+	+	+	+			+	+	+	+				
Shew.oneid.	3	+	+	+	+	+	+	+	+			+	+	+		+	+		
Geob.metall.	3	+		+	+	+	+	+	+			+	+	+		+	+		
Hom.sapiens	1	+		+	+	+		+	+			+	+	+	+	+	+	+	+
Staph.aureus	-1																	+	+
E. coli K12	-1	+					+												

D



The development of this technology ensured that the subsystems information could be shared in a platform-independent manner, without requiring any centralized resource (such as a pathway collection). Any annotation environment can be developed or modified to support the creation and curation of subsystems using the clearinghouse (or, a local clearinghouse, if desired) as a repository.

THE SEED TECHNOLOGY TO SUPPORT SUBSYSTEMS

The SEED annotation environment is the first annotation environment that supports the creation, curation, population and exchange of subsystems. It supports publishing subsystems to a clearinghouse, and the downloading and installation of subsystems developed at other sites.

The SEED was developed by an international collaboration led by members of FIG and Argonne National Laboratory (6). The software is being made available as open source software released under the GNU public license (GPL) from the ftp site ftp://ftp.theseed.org/SEED.

Only a few enhancements would have to be added to any existing annotation system to support analysis of subsystems, and this functionality would extend existing software. The software would have to be extended to encode populated subsystems as objects and decode the populated subsystems as they are retrieved from the clearinghouse. Software would need to be included to publish and request populated subsystems from the clearinghouse. The software would have to be able to define the functional roles in initial subsystems, and to establish the subsystem connections between protein-encoding genes, functional roles and subsystems.

EXAMPLE POPULATED SUBSYSTEMS

Our populated subsystems were assembled into a single collection with a consistent formulation of functional roles and released via the web (http://www.theseed.org/Release1_Subsystems/index.html). An open source collection of software tools has been released via FTP ftp://ftp.theseed.org/SEED. To illustrate the advantages of subsystem based annotations over 'traditional' annotation systems several subsystems are described below:

Leucine Degradation and HMG-CoA Metabolism (http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=Leucine_Degradation_and_HMG-CoA_Metabolism&request=show_ssa)

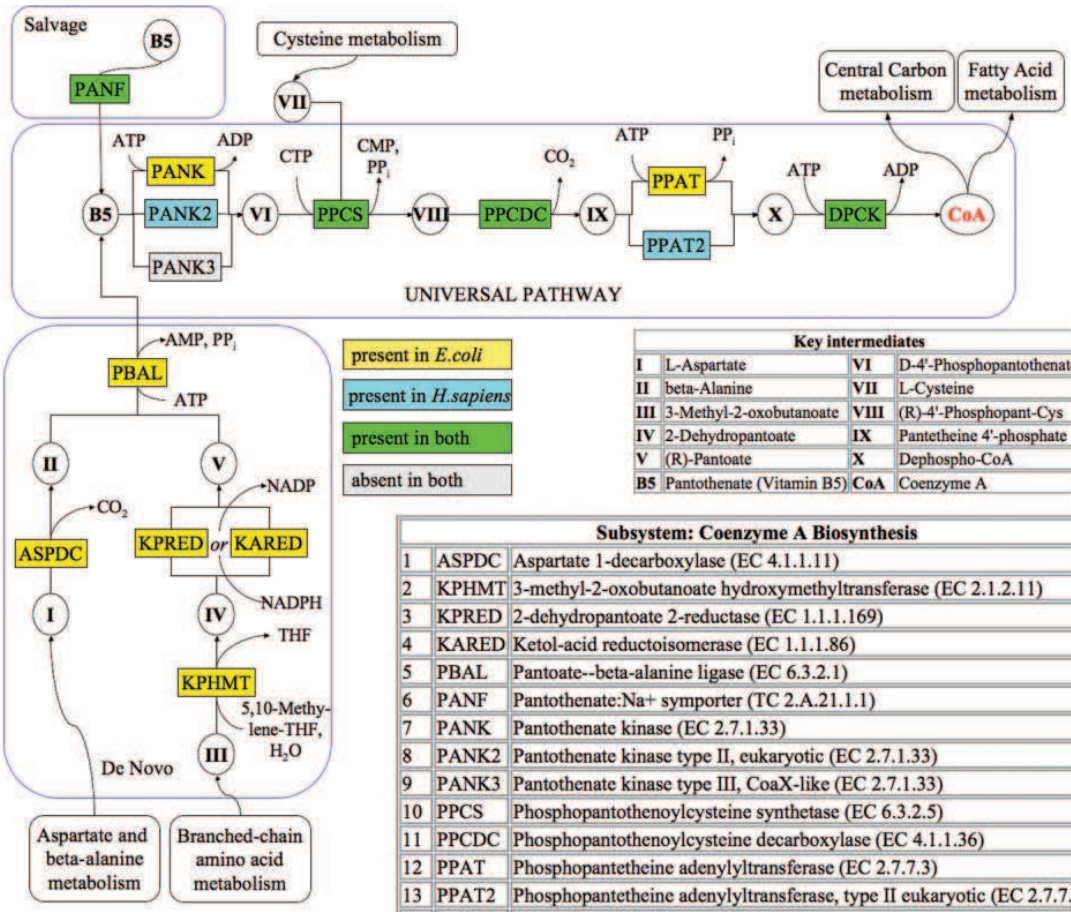
The populated subsystem presenting the leucine catabolism/HMG-CoA synthesis is depicted in Figure 3. An earlier analysis of some parts of this subsystem was presented elsewhere (7).

In humans leucine catabolism is coupled to sterol biosynthesis via a hydroxymethylglutaryl-coenzyme A (HMG-CoA) intermediate. The pathway is well characterized because defects in individual steps cause hereditary metabolic disorders like isovaleric acidemia, methylcrotonylglycinuria, methylglutaconic aciduria and 3-hydroxy-3-methylglutaric aciduria (8,9,10). Moreover, the human enzyme HMG-CoA reductase is a target in cardiovascular disease therapy because of its rate-limiting role in sterol biosynthesis (11). In contrast, only the early catabolic steps had been characterized in bacterial genomes—no genes were directly connected to enzymatic steps beyond isovaleryl-CoA (metabolite II in Figure 3B). Attempts to project from known eukaryotic genes based exclusively on homology searches produced ambiguous results because most of the enzymes in this pathway are members of large families of paralogs.

A combination of functional and genome context analysis, as depicted in the populated subsystem spreadsheet (Figure 3C) provided convincing evidence for the presence of the entire pathway of leucine catabolism in a number of diverse bacteria. A large conserved gene cluster containing reliable bacterial orthologs of two known human genes committed to this pathway was observed (Figure 3D). The gene *yngH* present in *Bacillus* and other bacteria is an ortholog of the human Methylcrotonyl-CoA carboxylase carboxyl transferase subunit (EC 6.4.1.4) while the neighboring gene *yngG* is an ortholog of HMG-CoA lyase (EC 4.1.3.4). This observation enabled the refinement of functional annotations for two additional bacterial genes in the same cluster (*yngJ*, an ortholog of Isovaleryl-CoA dehydrogenase (EC 1.3.99.10) and *yngF*, an ortholog of Methylcrotonyl-CoA carboxylase biotin-containing subunit (EC 6.4.1.4). Because these were weak homologs they could not be accurately characterized without considering the chromosomal neighborhood. The prediction (neither the bacterial nor the eukaryotic versions of methylglutaconyl-CoA hydratase were sequenced at that point) of *yng*G performing this function was projected from *Bacillus* to the human homolog. Later this prediction was proven correct by two independent publications that provided the experimental verification of the function encoded by this human gene (12,13).

Another functional inference from the analysis of this subsystem was a connection between leucine catabolism and acetoacetate metabolism (as illustrated in Figure 3B). This observation suggested a physiologically relevant extension of the HMG-CoA subsystem beyond its traditional boundaries. Two forms of *yngF* (encoding the methylcrotonyl-CoA carboxylase biotin-containing subunit (EC 6.4.1.4) were observed—the most common form, a fusion of biotin carboxylase and a C-terminal biotin carboxylase carrier protein domain and a rare form, in which the biotin carboxylase and the downstream biotin carboxylase carrier protein-encoding gene are separate (as in *B.subtilis*). The subsystems

Figure 3. Leucine Degradation and HMG-CoA Metabolism Subsystem. Functional roles, abbreviations, key intermediates and reactions in the pathway diagram are presented using the same conventions as in Figure 2. (**A**) Functional roles in the subsystem. (**B**) The Subsystem diagram shows the presence of genes assigned with respective functions for *B.melitensis* and *G.metallireducens*, using color-coded highlighting as explained in the panel. (**C**) Subsystem spreadsheet showing presence of genes with functions is shown by gene names for *B.subtilis* or by '+' for all other genomes (modified from a regular SEED display showing all gene IDs). Highlighting by a matching color indicates proximity on the chromosome. (**D**) Clustering on the chromosome of genes involved in the Subsystem (large yellow cluster) demonstrated by alignment of the chromosomal contigs of respective genomes around a signature pathway gene, yngG. Homologous genes are shown by arrows with matching colors and numbers corresponding to functional roles in panel A. *B.subtilis genes* are marked by gene names. Other genes (not conserved within the cluster) are colored gray.



Key intermediates			
I	L-Aspartate	VI	D-4'-Phosphopantothenate
II	beta-Alanine	VII	L-Cysteine
III	3-Methyl-2-oxobutanoate	VIII	(R)-4'-Phosphopant-Cys
IV	2-Dehydropantoate	IX	Pantetheine 4'-phosphate
V	(R)-Pantoate	X	Dephospho-CoA
B5	Pantothenate (Vitamin B5)	CoA	Coenzyme A

Subsystem: Coenzyme A Biosynthesis		
1	ASPDC	Aspartate 1-decarboxylase (EC 4.1.1.11)
2	KPHMT	3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11)
3	KPRED	2-dehydropantoate 2-reductase (EC 1.1.1.169)
4	KARED	Ketol-acid reductoisomerase (EC 1.1.1.86)
5	PBAL	Pantoate--beta-alanine ligase (EC 6.3.2.1)
6	PANF	Pantothenate:Na+ symporter (TC 2.A.21.1.1)
7	PANK	Pantothenate kinase (EC 2.7.1.33)
8	PANK2	Pantothenate kinase type II, eukaryotic (EC 2.7.1.33)
9	PANK3	Pantothenate kinase type III, CoaX-like (EC 2.7.1.33)
10	PPCS	Phosphopantothenoylcysteine synthetase (EC 6.3.2.5)
11	PPCDC	Phosphopantothenoylcysteine decarboxylase (EC 4.1.1.36)
12	PPAT	Phosphopantetheine adenylyltransferase (EC 2.7.7.3)
13	PPAT2	Phosphopantetheine adenylyltransferase, type II eukaryotic (EC 2.7.7.3)
14	DPCK	Dephospho-CoA kinase (EC 2.7.1.24)

Examples/ number of genomes	variant	ASPDC	KPHMT	alternatives		PBAL	PANF	alternatives			PPCS	PPCDC	alternatives		DPCK
				KPRED	KARED			PANK	PANK3	PANK2			PPAT	PPAT2	
De novo, complete /100		+	+	+ or/and +		+	±	+or/and + or +			+	+	+		+
Esc.coli K12	1a	*panD*	*panB*	*panE*	*ilvC*	*panC*	*panF*	*coaA*			*coaBC*	*coaBC*	*coaD*		*coaE*
Dein.radiod..	1b	900	2793		1701	1347	659		643		761	761	824		2074
Staph.aur.	1c	2477	2479	2480	1926	2478	1771+			1999	1089	1089	1004		1561
De novo (ASPDC=?)/45		?	+	+ or/and +		+	±	+or/and + or +			+	+	+ or +		+
Shew. onei.	2a	?	810	3466	3948	809		199			3851	3851	4255		389
Geob.metalli.	2b	?	179	2031	2898	178	1470		1816		2004	2004	67		1465
Sach. cerev.	2c	?	377	755	3888	2637				1397	2697	3297+		2314	1076
De novo, archaea/8		?	+	+ or/and +		?	±	?			+	+		+	?
Pyrob.aeroph.	3	?	2402	2401		?		?			831	831		608	?
Salvage of B5/40							+/?	+or/and + or +			+	+	+ or +		+
Strep.pneum	4a				403			741			1109	1110	1781		873
Therm.teng.	4b				15				2203		1410	1410	1388		814
Hom.sap	4c						11370+			10185+	12518	11601		12914	12914-
Truncated pathways/20		±	±	±		±		±					±		+
Buch.aphid	5	?	196		566	195							554		202
Trep.pallid.	6								430				283		296
Chlam.trach.	7														504

approach allows for different variants of enzymes as shown in Figure 3.

Panels B and C in Figure 3 illustrates the analysis of *functional variants* of a subsystem. Most of the subsystem protein-encoding genes are conserved in those species that have a functional ('nonzero') variant. However, *E.coli* and *Staphylococcus aureus* do not have a functional variant leading to the inference that they are incapable of catabolizing leucine using this pathway. Consequently, they were marked '−1' in the subsystem spreadsheet (Figure 3C). A distinction between the functional variants 1–3 was made based on the downstream component of the subsystem: the alternative routes of conversion of acetoactetate to succinate (intermediate V in Figure 3B). This was either via Succinyl-CoA:3-ketoacid-coenzyme A transferase subunits A and B (EC 2.8.3.5) (variant 2; e.g. *Brucella melitensis*) or via Acetoacetyl-CoA synthetase (EC 6.2.1.16) (variant 3; e.g. *Geobacter metallireducens* and *Shewanella oneidensis*). Both routes were possible in variant 1, as exemplified by both human and *B.subtilis*, although clustering on the chromosome suggests that in the latter species an AACS-dependent reaction may be preferred or co-regulated with the other components of the subsystem.

This example illustrates how prokaryotic chromosomal clustering can influence the interpretation of pathways, prediction of missing genes and projection of annotations between prokaryotic and eukaryotic genes. The observations also contributed to interpretation of the evolutionary history of a large and diversified group of proteins. More such examples have been published elsewhere (3,14).

Coenzyme A biosynthesis subsystem (http://www.theseed.org/annocopy/FIG/subsys.cgi?ssa_name=Coenzyme_A_Biosynthesis&request=show_ssa)

Coenzyme A (CoA) is a universal and essential cofactor in all forms of cellular life (15). Earlier bioinformatics analysis of CoA biosynthesis revealed a number of interesting variations between species (3,16,17). In the respective SEED subsystem (see Figure 4), this analysis was extended to >250 diverse genomes. A five-step pathway from pantothenate (vitamin B_5) to CoA is the universal component of the subsystem conserved in the majority of species. The most variable aspect of this pathway is pantothenate kinase (PANK). Three non-orthologous forms of PANK are presently known, and, in some cases, two alternative forms are present in the same organism. A recently identified and characterized CoaX-like (type III) pantothenate kinase (PANK3) appears to be more common in the bacterial world than the 'classic' PANK1 (18). Nevertheless, in most genomes, homologs of PANK3 have misleading annotations (e.g. 'BVG accessory factor'). The populated subsystem allows one to suggest reliable annotations for these proteins in many bacterial genomes, strongly supported by the strict requirement of PANK for CoA biosynthesis. The eukaryotic-like PANK2 was predicted (19) and subsequently verified (20) as the only PANK in all *Staphylococcus* species.

A possible fourth non-orthologous form of PANK can be inferred from the analysis of Archaea. The candidate for the missing archaeal PANK is a member of the GHMP kinase family which clusteres on the chromosome with several other CoA biosynthetic genes in some Archaea (i.e. PAE3407 of *Pyrobaculum aerophilum*). Another conserved family (represented by PAE1629 of *P.aerophilum*) may fulfill the role of dephospho-CoA kinase (DPCK), which is still 'missing' in all Archaea. This conjecture is based on a long-range sequence similarity with bacterial and eukaryotic enzymes (as suggested by the tentative annotation of COG0237 at NCBI http://www.ncbi.nlm.nih.gov/COG/old/palox.cgi?COG0237).

Both functional predictions [also suggested by (17)] require experimental verification. Among other problems within this subsystem is a missing aspartate decarboxylase in a number of genomes with an otherwise complete set of genes for the *de novo* synthesis.

Several examples illustrating major functional variants of the subsystem are outlined in Figure 4. An algorithm of semi-automated variant classification and a brief analysis of the key operational variants of CoA biosynthesis were recently published (21). Most species implement either complete *de novo* biosynthesis (variants 1–3) or a five-step pantothenate salvage (variant 4). A relatively small group of bacteria, most notably obligate intracellular pathogens and symbionts, display a variety of truncated pathways. For example, a disrupted pattern (missing PANK, PPCS and PPCDC) observed in *Buchnera aphidicola* suggests a possible *metabolic exchange* between this endosymbiont and the aphid host cell. According to this hypothesis, pantothenate produced but not utilized by *B.aphidicola* may be fed directly into the universal pathway of the host. The latter may *pay back* by providing a phosphopantetheine intermediate required for the last two steps of CoA synthesis in *B.aphidicola*. Several other interesting aspects of this subsystem are discussed in the supplementary materials (http://www.theseed.org/Release1_Subsystems/index.html).

Figure 4. CoA Biosynthesis Subsystem. Functional roles, abbreviations, key intermediates and reactions in the pathway diagram are presented using the same conventions as in Figure 2. Background colors in the diagram illustrate the comparison of subsystem variants by highlighting functional roles asserted in two organisms: *E.coli* (yellow) and *H.sapiens* (blue). Shared functional roles are highlighted green. The lower panel is a modification of the subsystem spreadsheet. It shows a classification of major subsystem variants representing a substantially different reaction topology revealed by semi-automated graph analysis as described in (21). Selected genomes unambiguously associated with each variant are shown after variant description (e.g. *De novo*, complete/100). Patterns of functional roles which constitute each functional variant are generalized by: '+', presence of a gene (for a given role) is required; '±', optional; '?', function is inferred by pathway analysis but a gene is unknown or 'missing' (i.e. can not be located by similarity). Typical sub-variants characterized by the same topology but relying on alternative (non-orthologous) forms of specific enzymes (e.g. PANK) are illustrated by the following genomes: *E.coli* K12 [NCBI taxonomy ID 83333.1], *D.radiodurans* R1 [243230.1], *S.aureus* subsp. *aureus* N315 [158879.1], *S.oneidensis* MR-1 [211586.1], *G.metallireducens* [28232.1], *Saccharomyces cerevisiae* [4932.1], *P.aerophilum* str. IM2 [178306.1], *Streptococcus pneumoniae* R6 [171101.1], *Thermoanaerobacter tengcongensis* [119072.1], *H.sapiens* [9606.2], *B.aphidicola* str. APS (*Acyrthosiphon pisum*) [107806.1], *Treponema pallidum* subsp. *pallidum* str. Nichols [243276.1] and *Chlamydia trachomatis* D/UW-3/CX [272561.1]. Genes assigned with respective functional roles are shown by SEED unique IDs for all illustrated genomes (except *E.coli* where common gene names are used). Matching background colors highlight genes that occur close to each other on the chromosome.

Ribosomal proteins (http://www.theseed.org/SubsystemStories/Ribosomal_proteins/abstract.htm)

Historically, ribosomal proteins were identified in several important experimental organisms, including *E.coli*, *Bacillus* species, yeast, rat and *Halobacterium.* In each case, a unique nomenclature was developed. More recently, several groups sought unified nomenclatures given the availability of so many sequences. In the cases of Bacteria and Eukarya, these efforts were hugely successful. The most problematic aspects of the conventions were (i) the failure to uniformly indicate whether a given label is based upon the bacterial or the eukaryal numbering, and (ii) the linking of equivalent eukaryal and bacterial terms. There are only two proteins (S3 and L3) for which the bacterial and eukaryal numbers are the same. This created a particularly confusing situation when the bacterial nomenclature was applied to Archaea, except when no bacterial homolog existed, in which case the eukaryal label was applied.

To address these problems a dual labeling was applied in which bacterial proteins were given the bacterial label (always explicitly including the 'p', e.g. S5p), followed by the designation of the corresponding eukaryal protein in parentheses (always with the explicit 'e', e.g. S2e). Similarly, in the case of eukarya, the eukaryal protein designation is given first, followed by the bacterial label in parentheses. In the case of Archaea, in all but a few cases the proteins are clearly of the eukaryal genre, and the eukaryal term is given first. One of the most important consequences of this nomenclature is that a text-based search is always unambiguous as to whether the bacterial or eukaryal numbering is desired. For example, a search for L11p will return bacterial L11 and eukaryal L12, but not bacterial L5 (the equivalent of eukaryal L11). A second key decision was to use the terms LSU and SSU to distinguish the subunits, rather than 30S, 40S, 50S and 60S. In addition to further unifying the nomenclature, it avoids two key sources of confusion. Several eukaryal ribosomes (especially organellar ribosomes) have been assigned to 'non-standard' sizes. Thus, searching for 50S and/or 60S was not sufficient to ensure that all ribosomes were distinguished. But more importantly, it avoids the temptation to use 50S to designate the LSU of a eukaryal mitochondrial ribosome. Instead, we have explicitly identified all organellar proteins by 'mitochondrial' or 'chloroplast'.

The development of this nomenclature demonstrated the power of the subsystems approach for encoding non-metabolic pathways, and the utility of functional roles in describing a controlled vocabulary for gene product function.

THE IMPACT OF POPULATED SUBSYSTEMS

As demonstrated by the examples above, populated subsystems can be used to support two broad categories of research: advancing research in the populated subsystems themselves and addressing numerous fundamental problems within bioinformatics.

It is important to note that there are large and ongoing efforts that address similar objectives—most notably the KEGG (http://www.genome.jp/kegg/kegg2.html) (22,23), GO (http://www.geneontology.org/) (5) and MetaCyc (http://metacyc.org/) (24) projects. These represent substantial projects, and we have in many ways built upon their work. Perhaps, the most obvious difference between our work and these projects is that we have made it possible for all researchers to immediately develop detailed encodings of their particular area of expertise, to make these new encodings available to the research community, and to import the work of others in constructing a customized collection of subsystems covering their specific needs. This radically decentralized effort offers a different set of incentives for domain experts to participate, which is precisely what will be needed to improve existing annotations.

The primary utility of annotated subsystems relates to the fact that a populated subsystem often supports substantially more accurate assignments of function to genes.

In addition the analysis of the populated subsystem allows one to arrive at a precise notion of which forms (i.e. which variants) of the subsystem exist in which organisms.

Further, the spreadsheet included in an populated subsystem often makes it vividly clear that a gene implementing a specific functional role is very likely to exist, even though it has not yet been identified. These so-called *missing gene* problems occur with surprising frequency. In the two metabolic examples presented in this paper and in various instances published in the Supplemental Material we show in detail a few instances in which conjectures could easily be formulated once the actual presence of a missing gene had been identified.

Finally, the presence of an extensive set of annotated subsystems lays the foundation for an accurate characterization of the metabolic network present in each organism.

The existence of a collection of populated subsystems also has an impact on a number of important topics in bioinformatics:

(i) Over and over as we performed our analysis we found that genes that appeared to actually be missing in an annotated subsystem were, in fact, present within an open reading frame (ORF), but eluded identification by a gene-calling algorithm. For the functional roles represented in populated subsystems, it becomes possible to directly search for instances of these roles in cases in which there is reason to believe that such a gene must exist.

(ii) Once ORFs containing genes have been identified, the problem of accurately identifying the start of the gene remains. The most successful attempts have been based on alignments. We argue that use of genes that are both similar and believed to implement the same functional role will lead to substantial improvements over existing estimates. A team at Middle Tennessee State University has brought up a website (http://torvalds.cs.mtsu.edu/cgi-bin/starts/starts.cgi) with initial results.

(iii) The search for regulatory sites in upstream regions of related genes has often led to success (25). Regulon analysis in combination with other techniques of comparative genomics was allowed to improve interpretation and to generate functional predictions in a number of metabolic subsystems (26,27). With the release of our initial set of annotated subsystems, we are making data available to support such analysis. For each annotated subsystem, we are providing sequences of upstream regions for each prokaryotic genome. Each sequence contains 300 bp of upstream sequence depicting the boundary of the adjacent gene (delimiting the intergenic gap), as well as 100 bp of the gene sequence itself.

(iv) The development of carefully curated protein families has historically been a key goal of bioinformatics for obvious reasons. The limitations of existing formulations relate to ambiguities in function assignment, a problem that is directly addressed by annotated subsystems. We have used this initial collection to create a list of refinements of UniProt annotations, and we will work to make sure that our analysis directly supports both the UniProt and other efforts to produce clean, comprehensive collections of protein families.

(v) Some of the most successful applications of bioinformatics technology relate to *context analysis* (3,28,29). In numerous cases, the clues that led to conjectures of function were based on the fact that related genes tend to cluster on prokaryotic chromosomes, tend to fuse and to co-occur. The annotated subsystems offer a framework for establishing the statistical properties needed to effectively exploit these tendencies.

(vi) The long-term goal of the subsystems approach is to bring every subsystem to a point where it has been carefully curated by one or more experts in the biological process encoded by the given subsystem. This approach will lead to the construction of an accurate phylogenetic context for each of the proteins within the subsystem, resulting in the ability to accurately trace the evolutionary histories of the catalytic domains that make up each subsystem [for a detailed illustration of this style of analysis, see ref. (30)].

(vii) Subsystems have also provided an approach for understanding the metabolism of environmental samples. A comparison of statistically significantly different subsystems present in different large environmental (metagenome) samples yielded unprecedented insights into the biology of these environments and lead to the generation of novel hypotheses to be tested by field biologists (R. Edwards, unpublished data).

THE RELEASE

Concurrent with the publication of this paper, an initial snapshot release of our collection of populated subsystems (which was a subset of those available via the SEED clearinghouse) was made. This subset is available in a format that makes the data easily accessible for use in other systems or as raw data. The current release of 173 populated subsystems is available without restriction via the web. The supplementary online subsystems material includes three main components:

(i) A set of 48 example subsystems. These constitute examples that have been curated in somewhat more detail and have led to interesting conjectures or research results in a number of cases. For each of these examples, we include the complete subsystem 'frozen' at the time of release, abstracts, presentations or summaries providing more detail about the subsystem and suitable for classroom use or lectures.

(ii) A set of 173 populated subsystems 'frozen' at the time of release that cover a large swath of central metabolism and other cellular processes.

(iii) Links to the current status of each subsystem. Each of the subsystems is continually being curated and populated as new genomes are added to the SEED and new comparisons become available. These links provide access to the most up-to-date annotations.

Each provided sequence was packaged with as many IDs as possible. For example, identifiers from FIG, UniProt, KEGG and NCBI (including GI number, gene number, UI or RefSeq ID), as well as identifiers from sequencing laboratories were included to ensure portability. The SEED release is itself open source software and can be acquired via FTP ftp://ftp.theseed.org/SEED. The system was developed to run on both Mac OSX systems and Linux systems.

CONCLUSIONS

Within 2–3 years we will all have access to over a thousand sequenced genomes. This data will grow to become the central resource in modern biology. Annotating this collection is the core challenge of modern bioinformatics. In this paper we describe a new approach to annotation based on idea of subsystems that promises to dramatically improve the quality and utility of annotations. This approach is central to the Project to Annotate 1000 genomes and has been implemented in a suite of tools for genome annotation. The approach and technology provide one way to involve many domain experts in the genome annotation process. The technology for developing these subsystems now exists, the technologies for supporting automated addition of new genomes to the collection of populated subsystems is now being developed, and the initial collection is being made available to the research community.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the Fellowship for Interpretation of Genomes.

Conflict of interest statement. None declared.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. Haft,D.H., Selengut,J.D., Brinkac,L.M., Zafar,N. and White,O. (2005) Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics*, **21**, 293–306.
3. Osterman,A. and Overbeek,R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.*, **7**, 238–251.
4. Overbeek,R., Larsen,N., Smith,W., Maltsev,N. and Selkov,E. (1997) Representation of function: the next step. *Gene*, **191**, GC1–GC9.
5. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genet*, **25**, 25–29.
6. Overbeek,R., Disz,T. and Stevens,R. (2004) The SEED: a peer-to-peer environment for genome annotation. *Commun. ACM*, **47**, 46–51.
7. Overbeek,R., Devine,D. and Vonstein,V. (2003) Curation is forever: comparative genomics approaches to functional annotation. *Targets*, **2**, 138–146.

8. Tanaka,K., Ikeda,Y., Matsubara,Y. and Hyman,D.B. (1987) Molecular basis of isovaleric acidemia and medium-chain acyl-CoA dehydrogenase deficiency. *Enzyme*, **38**, 91–107.
9. Weyler,W., Sweetman,L., Maggio,D.C. and Nyhan,W.L. (1977) Deficiency of propionyl-Co A carboxylase and methylcrotonyl-Co A carboxylase in a patient with methylcrotonylglycinuria. *Clin. Chim. Acta.*, **76**, 321–328.
10. Gibson,K.M., Lee,C.F. and Hoffmann,G.F. (1994) Screening for defects of branched-chain amino acid metabolism. *Eur. J. Pediatr.*, **153**, S62–67.
11. Marz,W. and Wieland,H. (2000) HMG-CoA reductase inhibition: anti-inflammatory effects beyond lipid lowering? *Herz*, **25**, 117–125.
12. Loupatty,F.J., Ruiter,J.P., L,I.J.1st, Duran,M. and Wanders,R.J. (2004) Direct nonisotopic assay of 3-methylglutaconyl-CoA hydratase in cultured human skin fibroblasts to specifically identify patients with 3-methylglutaconic aciduria type I. *Clin. Chem.*, **50**, 1447–1450.
13. Ly,T.B., Peters,V., Gibson,K.M., Liesert,M., Buckel,W., Wilcken,B., Carpenter,K., Ensenauer,R., Hoffmann,G.F., Mack,M. *et al.* (2003) Mutations in the AUH gene cause 3-methylglutaconic aciduria type I. *Hum. Mutat.*, **21**, 401–407.
14. Jordan,I.K., Henze,K., Fedorova,N.D., Koonin,E.V. and Galperin,M.Y. (2003) Phylogenomic analysis of the Giardia intestinalis transcarboxylase reveals multiple instances of domain fusion and fission in the evolution of biotin-dependent enzymes. *J Mol. Microbiol. Biotechnol.*, **5**, 172–189.
15. Begley,T.P., Kinsland,C. and Strauss,E. (2001) The biosynthesis of coenzyme A in bacteria. *Vitam. Horm.*, **61**, 157–171.
16. Gerdes,S.Y., Scholle,M.D., D'Souza,M., Bernal,A., Baev,M.V., Farrell,M., Kurnasov,O.V., Daugherty,M.D., Mseeh,F., Polanuyer,B.M. *et al.* (2002) From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. *J. Bacteriol.*, **184**, 4555–4572.
17. Genschel,U. (2004) Coenzyme A biosynthesis: reconstruction of the pathway in archaea and an evolutionary scenario based on comparative genomics. *Mol. Biol. Evol.*, **21**, 1242–1251.
18. Brand,L.A. and Strauss,E. (2005) Characterization of a new pantothenate kinase isoform from Helicobacter pylori. *J. Biol. Chem.*, **280**, 20185–20188.
19. Daugherty,M., Polanuyer,B., Farrell,M., Scholle,M., Lykidis,A., de Crecy-Lagard,V. and Osterman,A. (2002) Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *J. Biol. Chem.*, **277**, 21431–21439.
20. Choudhry,A.E., Mandichak,T.L., Broskey,J.P., Egolf,R.W., Kinsland,C., Begley,T.P., Seefeld,M.A., Ku,T.W., Brown,J.R., Zalacain,M. *et al.* (2003) Inhibitors of pantothenate kinase: novel antibiotics for staphylococcal infections. *Antimicrob. Agents Chemother.*, **47**, 2051–2055.
21. Ye,Y., Osterman,A., Overbeek,R. and Godzik,A. (2005) Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics*, **21**, 478–486.
22. Kanehisa,M. (1997) A database for post-genome analysis. *Trends Genet*, **13**, 375–376.
23. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
24. Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–442.
25. Gelfand,M.S., Novichkov,P.S., Novichkova,E.S. and Mironov,A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Brief Bioinform.*, **1**, 357–371.
26. Rodionov,D.A., Vitreschak,A.G., Mironov,A.A. and Gelfand,M.S. (2002) Comparative genomics of thiamin biosynthesis in procaryotes: new genes and regulatory mechanisms. *J. Biol. Chem.*, **277**, 48949–48959.
27. Rodionov,D.A., Mironov,A.A. and Gelfand,M.S. (2002) Conservation of the biotin regulon and the BirA regulatory signal in eubacteria and archaea. *Genome. Res.*, **12**, 1507–1516.
28. Koonin,E.V. and Galperin,M.Y. (2002) *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*. 1st Edn Kluwer Academic Publishers, Boston.
29. Huynen,M.A., Snel,B., von Mering,C. and Bork,P. (2003) Function prediction and protein networks. *Curr. Opin. Cell. Biol.*, **15**, 191–198.
30. Xie,G., Keyhani,N.O., Bonner,C.A. and Jensen,R.A. (2003) Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol. Mol. Biol. Rev.*, **67**, 303–342.

Paper IV

BIOINFORMATICS

Vol. 22 no. 14 2006, pages e281–e289
doi:10.1093/bioinformatics/btl247

Finding novel genes in bacterial communities isolated from the environment

Lutz Krause[1,*], Naryttza N. Diaz[1], Daniela Bartels[1], Robert A. Edwards[2,3,4], Alfred Pühler[6], Forest Rohwer[3,4], Folker Meyer[1] and Jens Stoye[5]

[1]Bielefeld University, Center for Biotechnology (CeBiTec) D-33594 Bielefeld, Germany, [2]Fellowship for Interpretation of Genomes, Burr Ridge IL, [3]Department of Biology, San Diego State University, San Diego, CA, [4]Center for Microbial Sciences, San Diego, CA, [5]Universität Bielefeld, Technische Fakultät D-33594 Bielefeld, Germany and [6]Universität Bielefeld, Lehrstuhl für Genetik, Fakultät für Biologie D-33594 Bielefeld, Germany

ABSTRACT

Motivation: Novel sequencing techniques can give access to organisms that are difficult to cultivate using conventional methods. When applied to environmental samples, the data generated has some drawbacks, e.g. short length of assembled contigs, in-frame stop codons and frame shifts. Unfortunately, current gene finders cannot circumvent these difficulties. At the same time, the automated prediction of genes is a prerequisite for the increasing amount of genomic sequences to ensure progress in metagenomics.

Results: We introduce a novel gene finding algorithm that incorporates features overcoming the short length of the assembled contigs from environmental data, in-frame stop codons as well as frame shifts contained in bacterial sequences. The results show that by searching for sequence similarities in an environmental sample our algorithm is capable of detecting a high fraction of its gene content, depending on the species composition and the overall size of the sample. The method is valuable for hunting novel unknown genes that may be specific for the habitat where the sample is taken. Finally, we show that our algorithm can even exploit the limited information contained in the short reads generated by 454 technology for the prediction of protein coding genes.

Availability: The program is freely available upon request.

Contact: Lutz.Krause@CeBiTec.Uni-Bielefeld.DE

1 INTRODUCTION

Novel sequencing methods have recently revolutionized the field of genome research. The sequencing of samples isolated directly from the environment allows access to organisms that can not be cultivated in the laboratory (Breitbart *et al*. (2002), Tyson *et al*. (2004), Venter *et al*. (2004)). Additionally, the massively parallel pyrosequencing system which was recently developed by 454 Life Science, Inc, has dramatically dropped the time and cost constraints of DNA sequencing (Margulies *et al*. (2005)). The application of 454 technology provides larger amounts of sequences at a lower cost compared to traditional DNA sequencing methods. These sequences are of great value for the identification of novel genes that can not be found in organisms cultured with traditional methods. The importance of such approaches is stressed by the fact that only a fraction of the living organism found in natural environments can be cultured by conventional methods (Tringe and Rubin (2005)).

The isolation and sequencing of DNA derived from diverse and mixed microbial communities is known as metagenomics, environmental genomics or ecogenomics. Although still in its infancy, this rapidly developing field has provided striking insights into the ecology and evolution of natural occurring microbial communities. Fields such as health and biotechnology have already benefited from metagenomics (Lombardot *et al*. (2006), Furrie (2006), Schloss and Handelsman (2003), Edwards and Rohwer *et al*. (2005), Edwards *et al*. (2006)).

Gene finding in environmental samples

Two different approaches are applied for predicting protein coding genes in bacterial genomes; intrinsic and extrinsic methods. Intrinsic methods (e.g. GLIMMER Delcher *et al*. (1999), GENEMARK Besemer and Borodovsky (1999)) analyze sequence properties of genomes to discriminate between coding sequences (CDS) and non-coding ORFs (NORFs). These methods exploit the different compositional properties of coding and non-coding sequences, which are mainly caused by a bias on codon usage in the CDS to optimize the translation efficiency (Gouy and Gautier (1982)).

In contrast, extrinsic methods (e.g. CRITICA Badger and Olsen (1999), ORPHEUS Frishman *et al*. (1998)) predict genes by searching for stretches of DNA that were conserved during evolution. The success of extrinsic methods can be explained by the fact that during evolution most of the new genes are formed by duplication, rearrangement and mutation events of existing genes (Chothia *et al*. (2003)).

The prediction of protein coding genes in environmental samples is problematic for several reasons. One is the low sequence quality of the assembled contigs which may lead to frame shifts and in-frame stop codons in the CDSs contained therein. Another problem is that assembled contigs may be too short to reveal the genome specific sequence properties, which are crucial in the application of intrinsic gene prediction methods. These reasons limit their application to environmental samples. Currently, the majority of the CDSs in environmental samples are identified based on a BLAST search against databases of known proteins.

*To whom correspondence should be addressed.

Species that are abundant in natural environments will also be over-represented in the samples. These species do not represent a problem while assembling them, and large stretches of their genomes can be obtained. But, the under-represented species constitute a challenge since for those only short contigs with low coverage are obtained. One problem related to the low coverage is that these contigs are even more prone to contain in-frame stop codons or frame shifts. Therefore, applying existing gene finders to environmental samples is fraught with difficulties because they were not designed to cope with this type of errors and short contigs.

Strategy

The main idea for the novel gene prediction method presented in this work is to search for stretches of DNA that are conserved within the environmental sample. Here, the algorithm does not rely on a pairwise sequence comparison, but instead it combines information from all BLAST hits at the same time. Conserved coding sequences are discriminated from conserved non-coding regions based on their synonymous substitution rate.

In functional proteins, the coding genes show a much higher number of synonymous substitutions than in non-coding sequences. The rate of synonymous to non-synonymous substitutions (k_S/k_A) reflects the interchange of positive selection and neutral evolution. Therefore, investigating the number of synonymous and non-synonymous substitutions can supply valuable information on whether or not a sequence stretch is under constraint for functional selection. This information can be used for the identification of genes in bacterial and eukaryotic genomes (Badger and Olsen (1999), Nekrutenko *et al*. (2003a), Nekrutenko *et al*. (2003b) and Moore and Lake (2003)).

For the prediction of protein coding sequences contained in a contig from an environmental sample, first a BLAST search against a nucleotide database is conducted. For this in principle any nucleotide database can be used, e.g databases containing complete genomes, metagenomes or known genes. To search for novel habitat specific genes a BLAST search against a database that exclusively contains all sequences from that sample can be employed. Subsequently, the algorithm needs to discriminate if the BLAST hits match conserved coding sequences, conserved non-coding regions, or shadows of CDSs in another reading frame. Additionally, a CDS may be embedded in long BLAST hits. For this case, the gene boundaries need to be identified. Given all BLAST hits for a contig, the algorithm will find the best path through all hits at the same time. In order to accomplish this task, several different features are taken into account: (a) the synonymous substitution rate at each position in the contig, (b) the positions of stop codons in the contig and (c) the position of stop codons in matching database sequences. Additionally, the end of BLAST hits are considered as possible indications for the boundaries of coding regions.

In the gene prediction process the algorithm will avoid in-frame stop codons, but otherwise will favor regions with a high synonymous substitution rate. The outcome of the BLAST hits are used to assign six scores to each nucleotide, one for each of the six possible reading frames, reflecting the nucleotides coding potential in this reading frame. Scores are assigned by counting the number of synonymous and non-synonymous substitutions at each position for each of the six reading frames. As a result, a scoring matrix with scores for each nucleotide in the contig is obtained. Based on these scores, a dynamic programming method is applied to find the optimal path through the matrix that maximizes the overall score (the sum of all scores on the path). The usage of a combined score for all BLAST hits should result in a superior performance compared to methods that rely on simple pairwise sequence alignments. The advantage should be particularly profound when a database of low quality with short contigs and many frame shifts is used for the BLAST based search for conserved sequences.

2 METHODS

The gene prediction algorithm

The algorithm can be divided into four phases: (1) a BLAST based search for conserved sequences (2) the calculation of combined scores (3) the prediction of coding sequences by dynamic programming and (4) the postprocessing.

Phase 1: Blast based search for conserved sequences During the first phase of the algorithm a BLAST search against a nucleotide database is conducted. Hereby, the contig as well as all sequences in the database are translated into all six reading frames (if the database contains known genes only the contig will be translated into all six reading frames). As the BLAST search is conducted on the amino acid level, each obtained hit is associated with a specific reading frame in the contig. The BLAST hits obtained are filtered, hits with $k_S/k_A < 1$ are excluded from the subsequent analysis as these do not indicate the presence of a coding sequence.

Phase 2: Calculation of combined scores In the second phase of the algorithm, the remaining hits are used to assess the coding potential of each nucleotide in the contig. Given a contig c of length n, $c[i]$ denotes the nucleotide at position i of that contig ($1 \leq i \leq n$). A nucleotide $c[i]$ could be coding in one of the six reading frames $k \in \{-3, -2, -1, +1, +2, +3\}$, or non-coding, denoted by $k = 0$. For each position i and for each reading frame k, the number of synonymous and non-synonymous substitutions at position i are counted (Figure 1). This is done by comparing the nucleotide sequence of the contig to the nucleotide sequence of all BLAST hits in this reading frame. The number of synonymous and non-synonymous substitutions are used to score that $c[i]$ is coding in reading frame k. Synonymous substitutions contribute with a positive score, non-synonymous substitutions with a negative score. Additionally, the correct ends of the coding sequences need to be determined. Therefore, stop codons in the contig are penalized with a negative score in the according frame. For a given BLAST hit both the contig and the matching database sequence of the BLAST hit may contain stop codons. To discriminate between real stop codons and stop codons introduced by sequencing errors, additionally negative scores are applied for: (a) all stop codons in the database sequences of the BLAST hits, (b) for ends of BLAST hits, as these also may indicate the boundaries of genes (Figure 1). Subsequently, each score obtained is normalized by the number of hits that contribute to that score. Using this strategy for all BLAST hits in reading frame k, a single combined score that reflects the coding potential of the contig at position i in this reading frame is derived. Additionally, for $k = 0$ a score of zero is assigned to each position i of the contig. As a result, a scoring matrix s_{ik} is derived which provides a position specific score that the contig is coding in one of the six reading frames or non-coding (Figure 1).

Phase 3: Prediction of coding sequences Coding sequences are predicted in the third phase. To assign one of the six reading frames k (or $k = 0$ for non-coding) to each position of the contig, the algorithm searches for the path in the scoring matrix s_{ik} that maximizes the sum of all scores on the path. According to the optimal path, each position i of the contig is subsequently labeled with the frame k it passes through at this position. Depending on their reading frame, genes may only start or

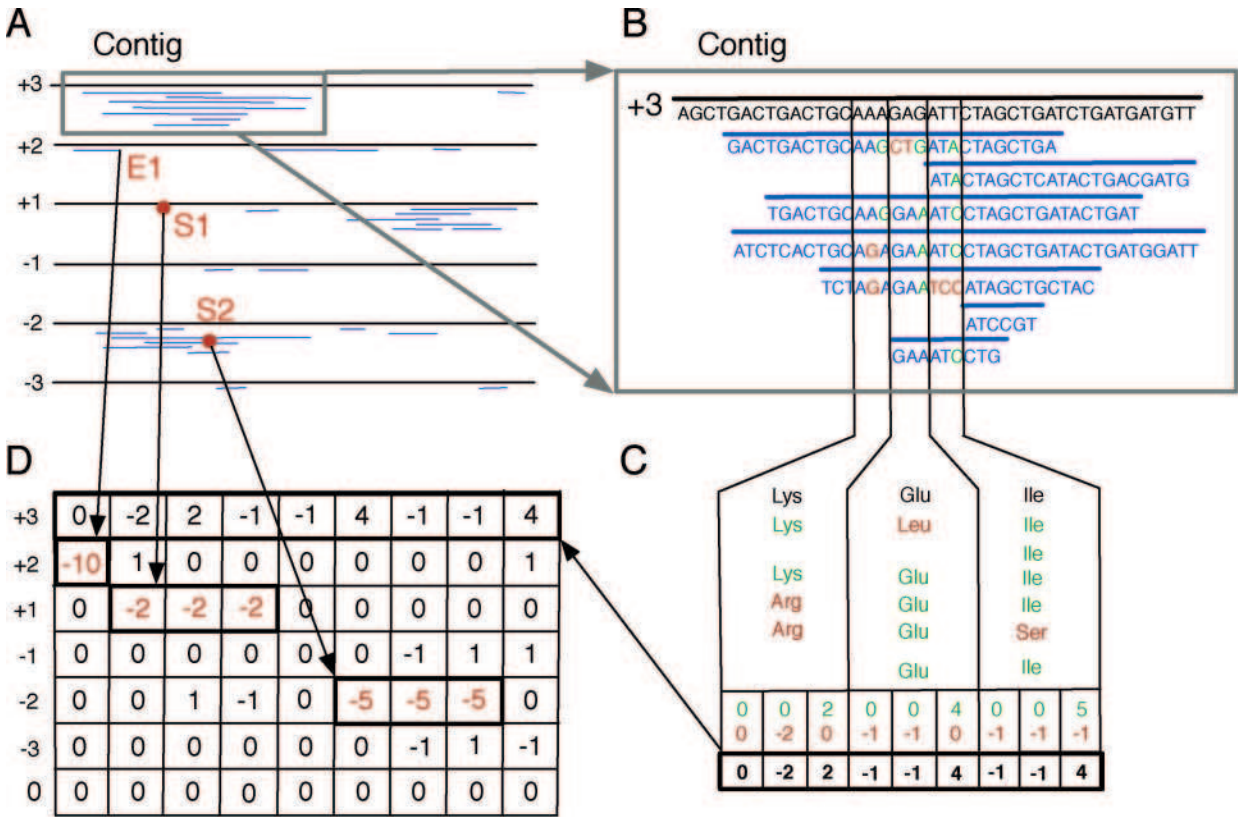


Fig. 1. Calculating combined scores. All scores are depicted without normalization. **A**) all six reading frames of a contig are shown (the continuous lines). BLAST hits matching the respective reading frames are displayed as blue bars below the reading frame. **B**) The nucleotide sequence of each reading frame of the contig is compared with all database sequences matching this reading frame. The number of synonymous and non-synonymous substitutions at each position is used as a score that the contig at this position is coding in the respective reading frame. **C**) The number of synonymous substitutions at each position are used as a positive score. The number of non-synonymous substitutions at each position contribute with a negative score. **D**) The calculated scores for each position and reading frame are stored in a matrix. For $k = 0$ a score of zero is assigned to each position i of the contig. Penalties are additionally added to the respective position and reading frame for stop codons in the contig (S1), in the matching database sequence (S2) as well as for the end of BLAST hits (E1).

stop at certain positions. Therefore, a valid path may not jump arbitrarily between frames, but instead underlies certain restrictions. To be precise, the set $V(i, k)$ of all valid precursors of a frame k at position i is defined as:

$$V(i,k) = \begin{cases} \{j, 0, -j\} & \text{if} \quad k = 0 \\ \{k, 0, -k\} & \text{if} \quad |k| = j \\ \{k\} & \text{otherwise.} \end{cases}$$

where $j = (i - 1) \bmod 3 + 1$. Figure 2 depicts the scoring matrix of combined scores and the calculation of the optimal path. This figure also introduces several terms used in the following. The optimal valid path for a scoring matrix s_{ik} can be calculated using dynamic programming by the following recursion:

$$f_i(k) = \max_{k' \in V(i,k)} \begin{cases} f_{i-1}(k') + s_{ik} + 2q & \text{if } k < 0 \text{ and } k' > 0 \\ f_{i-1}(k') + s_{ik} + q & \text{if } k \neq 0 \text{ and } k' \neq k \\ f_{i-1}(k') + s_{ik} & \text{otherwise} \end{cases}$$

where q is a negative score that is added to leave a gene on the forward strand or to enter a gene on the reverse strand ($2q$ are added if a gene on the forward strand is left and a gene on the reverse strand is entered at the same time). Thus, q is added for each 5′ end of a gene on a path. The penalty q was introduced to predict genes only in areas with sufficient coding evidence. The calculated value $f_i(k)$ is the maximal score of all paths that enter s at position 1 and pass through k at position i.

Phase 4: Postprocessing During the post-processing phase, the predictions are joined and frame shifts are identified. When a BLAST search against a database of short contigs is employed, genes may be covered only partially by hits which may result in the prediction of several fragments. This is particularly profound when a BLAST search against reads provided by the 454 technology is conducted. Therefore adjacent predictions within the same reading frame are joined if (a) their distance on the contig does not exceed 400 bp and (b) the sequence of the contig that separates the predictions does not contain an in-frame stop codon.

To identify frame shifts that were introduced by sequencing errors all adjacent predictions located on the same strand but within a different reading frame are predicted as frame shifts if (a) their distance on the contig is less than 200 bp and (b) they do do not have an in-frame stop codon close to the potential frame shift. As an optional postprocessing step, our algorithm can also extend predicted CDS to the longest possible ORF available for that prediction.

Implementation

The algorithm was implemented in PERL using an object oriented approach.

Measuring the performance

To evaluate the performance of the novel gene finder predictions were compared to known annotated genes. For this purpose, two measurements

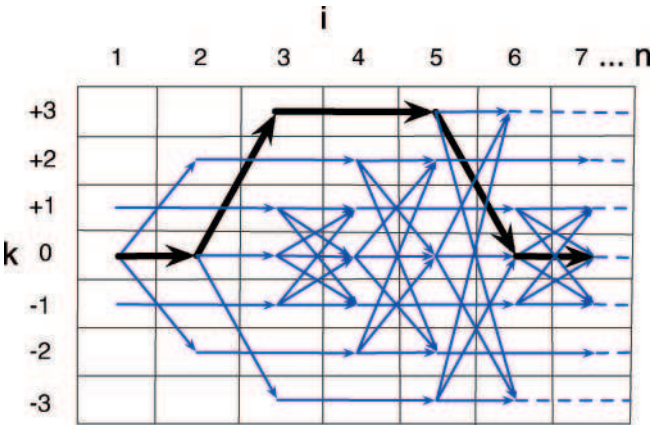


Fig. 2. Predicting coding sequences by calculating the optimal path in scoring matrix of combined scores. This figure shows the scoring matrix s_{ik} for the first seven positions of a contig. All valid paths in the scoring matrix are indicated with arrows. A gene is entered, if a path passes through frame $k \neq 0$ with the precursor frame $k' \neq k$. Accordingly, a gene is left, if a path that comes from a precursor frame $k' \neq 0$ enters a frame $k \neq k'$. The bold arrows depict an example path predicting a 3 bp long gene on the reading frame +3

are widely used: sensitivity and specificity. Sensitivity is a measure of the ability of the algorithm to predict known genes and is defined by $Sens = \frac{TP}{TP+FN}$. The specificity is a measure of the reliability of the predictions, given by the ratio $Spec = \frac{TP}{TP+FP}$. For the evaluation of the performance predictions were extended to the next 5′ stop codon. If an annotated CDS ends at that stop codon the prediction was counted as true positive (TP). Otherwise, the prediction was regarded as a false positive. All genes that are not completely embedded in the contigs are named truncated genes. These genes may appear at the end or beginning of the assembled contigs, therefore lacking the start or termination site of the gene. Truncated genes were excluded from the analysis.

Training GLIMMER on a synthetic metagenome

The prokaryotic gene finder GLIMMER version 3.01b was used to predict the genes of a synthetic metagenome (described in Materials). For the training step, all fragments of this metagenome were chained to one continuous contig. Adjacent fragments were concatenated with a linker sequence containing a stop codon in each of the six reading frames. Subsequently the GLIMMER ICM model was trained on the chained contig.

3 MATERIALS

Metagenome obtained with pyrosequencing

The performance of the algorithm was evaluated on a metagenome of a bacterial community isolated from the Solar Salterns in San Diego, CA (B. Rodriguez-Brito, R. Edwards, and F. Rohwer, Unpublished). Total community DNA was purified as described elsewhere (Edwards *et al*. (2006)) and sequenced using pyrosequencing by 454 Life Sciences, Inc, (Branford, CT). Using the 454 technology ≈60 Mb were obtained with an average read length of 100 bp. The reads were assembled using Phrap (Green (1994)). This resulted in 80,878 contigs with 16 Mb in total. In the following, this set is called *all contigs*. From this set a subset of contigs longer than 1,000 bp (2,244 contigs with 3.8 Mb in total) was selected, called hereafter *long contigs*.

Table 1. Annotated and published genomes used to create a synthetic metagenome

Organism	Accession number
Bacteria	
Alphaproteobacteria	
Candidatus pelagibacter ubique HTCC1062	NC_007205
Rhodobacter sphaeroides 2.4.1 chromosome 1	NC_007493
Gammaproteobacteria	
Shewanella oneidensis MR-1	NC_004347
Thiomicrospira crunogena XCL-2	NC_007520
Vibrio cholerae O1 biovar eltor str. N16961 chromosome 1	NC_002505
Cyanobacteria	
Prochlorococcus marinus subsp. *pastoris* str. CCMP1986	NC_005072
Synechococcus sp. WH 8120	NC_005070
Archaea	
Euryarchaeota	
Pyrococcus horikoshii OT3	NC_000961
Crenarchaeota	
Sulfolobus solfataricus P2	NC_002754

Species names and accessions numbers downloaded form the NCBI database.

The environmental sample from the Sargasso Sea

For the prediction of protein coding genes in metagenomes, the environmental sample from the Sargasso Sea (Venter *et al*. (2004)) was used as BLAST database during the search for conserved regions in the first phase of the algorithm. To save computational time, only half (≈390 Mb) of the entire Sargasso Sea sample was used.

Generating a synthetic metagenome

As a proof of concept, the algorithm was evaluated on a set of nine completely sequenced and annotated genomes (seven Bacteria and two Archaea, see Table 1). Members from the alpha, gammaproteobacteria and cyanobacteria groups were selected as they were reported to be abundant in the Sargasso Sea sample (Venter *et al*. (2004)). We also added Archaea to the evaluation set because they can be regarded as under-represented species in surface water marine environments. All genomic sequences and their respective annotations were downloaded from the NCBI Reference Sequence database (RefSeq) release 15 (Pruitt *et al*. (2005)). A synthetic metagenome with known CDSs was created by splitting the genome of each of the nine organism into fragments of length 4000 bp. A subset of *non-hypothetical* genes was created based on the annotated gene products from the public annotations. In this set all annotated genes with a gene product description of 'hypothetical protein' were excluded. Additionally, artificial sequencing errors (frame shifts and in-frame stop codons) were incorporated into all genes of the synthetic metagenome. In order to perform a systematic evaluation, all artificial sequencing errors were added to the synthetic metagenome in a controlled way. In one experiment, in-frame stop codons were added to the center of each gene of the original synthetic metagenome. In a second experiment frame shifts were incorporated to the center of all genes of the original synthetic metagenome.

4 RESULTS

Gene prediction in a synthetic metagenome using the environmental sample from the Sargasso Sea

Our algorithm can be used to identify genes contained in an environmental sample by directly searching for conserved regions within the sample. This approach may elucidate novel unknown genes present in the environmental sample which may be specific for the habitat the sample was taken from. The performance of the algorithm of predicting genes in an environmental sample by running a BLAST search against the sample itself was evaluated on the environmental sample data from the Sargasso Sea (Venter *et al*. (2004)). But, instead of drawing the contigs for which the genes are predicted directly from the Sargasso Sea sample, we used a more controlled and reliable data set. We chose several completely sequenced and annotated genomes from Bacteria groups that were also reported to be present in the species composition of the Sargasso Sea sample (Venter *et al*. (2004)). The genomes of these organisms were split into fragments of size 4000 bp, together forming a synthetic metagenome as a reliable standard of truth. Subsequently the genes of these contigs were predicted with our algorithm based on a BLAST search against the Sargasso Sea sample. To accurately evaluate the prediction performance that can be expected for a 'real' metagenome, sequences from alpha and gammaproteobacteria which are reported as over-represented in the Sargasso Sea sample, cyanobacteria which are modest abundant, as well as sequences from extremely scarce groups (two Archaea members) were included. The performance of the algorithm was measured by comparing the genes predicted for the synthetic metagenome with the known genes from the public genome annotations. To additionally evaluate the performance for sequencing errors that may frequently occur in metagenomes, three validation sets were used: (1) synthetic metagenome without artificial sequencing errors, (2) synthetic metagenome with in-frame stop codons and (3) synthetic metagenome with frame shifts.

Experiment 1: Gene prediction in a synthetic metagenome without artificial sequence errors The sensitivity and specificity reached by the algorithm for each organism contained in the synthetic metagenome is shown in Table 2. The results show that the sensitivity of the method strongly depends on the abundance of the different groups of Bacteria in the sample. While for the more abundant alpha, gammaproteobacteria and cyanobacteria an average sensitivity of 79% for all genes and 89% for the subset of non-hypothetical genes is achieved, for the Archaea the sensitivity is strongly reduced. The lower sensitivity for the Archaea was expected because this group is very rare in surface water marine environments and hence extremely scarce in the environmental sample from the Sargasso Sea. For the two cyanobacteria contained in the synthetic metagenome even a sensitivity of more than 94% is achieved for the non-hypothetical genes. In contrast, with a specificity between 88% and 99% the algorithm is highly specific for all groups. On average the specificity is 95%. The considerably lower overall sensitivity ($Sens_{all}$) when compared to the sensitivity for the subset of non-hypothetical genes ($Sens_{nh}$) can be explained by the fact that most of the genes labeled as 'hypothetical protein' in the public annotations were originally predicted with intrinsic methods. Many of these genes are either orphans (genes without sequence similarity to any known gene) or in fact non-coding and hence wrong annotations.

Table 2. Performance for a synthetic metagenome evaluated on the Sargasso Sea environmental sample

Organism	$Sens_{all}$	$Sens_{nh}$	Specificity
Bacteria			
Alphaproteobacteria			
C. pelagibacter	91.07	93.76	97.63
R. sphaeroides	62.01	77.62	97.02
Gammaproteobacteria			
S. oneidensis	85.36	95.12	90.44
T. crunogena	65.38	79.33	97.54
V. cholerae	69.66	87.66	93.88
Cyanobacteria			
P. marinus	93.29	94.42	89.75
Synechococcus sp.	82.99	94.13	87.83
Archaea			
Euryarchaeota			
P. horikoshii	29.99	66.40	97.77
Crenarchaeota			
S. solfataricus	26.69	43.44	98.89
Average	67.38	81.32	94.53

$Sens_{all}$ refers to the sensitivity calculated over all genes contained in the synthetic metagenome. $Sens_{nh}$ is the sensitivity calculated over all non-hypothetical genes. The entire Bacteria group represents the most common organisms in the Sargasso Sea sample. While the Archaea is the extremely scarce set for surface water marine environment.

Experiment 2: Gene prediction in a synthetic metagenome with artificial in-frame stop codons In the second experiment the performance of the algorithm was evaluated on genes containing in-frame stop codons. Therefore, an in-frame stop codon was added to the center of each annotated gene in the synthetic metagenome. In addition to the sensitivity and specificity, the percentage of true positives (TP) that span the artificially added stop codons was measured. In comparison to the synthetic metagenome without artificial sequence errors, for the genes with in-frame stop codons only a slight reduction in sensitivity and specificity was registered. The sensitivity is reduced by 1.7% for all genes and 1.3% for the subset of non-hypothetical genes. The reduction in specificity is 0.3%. On average, for 77% of all identified genes (TP) the prediction also spans the added in-frame stop codon (Table 3) and therefore correctly recognizes the stop codon as sequencing error. Strikingly, for the synthetic metagenome without artificial sequence errors only 4 predictions wrongly span a 'real' stop codon terminating the translation. These results demonstrate that the algorithm is quite robust for the task of identifying functional genes containing in-frame stop codons, generated by sequencing errors. These results also reveal the strength of our method to incorporate several features to determine the boundaries of coding sequence and to discriminate between 'real' stop codons and those introduced by sequencing errors.

Experiment 3: Gene prediction in a synthetic metagenome with artificial frame shifts In the third experiment the performance of the novel algorithm to predict frame shifts introduced by sequencing errors was evaluated. Therefore, an artificial frame shift was

Table 3. Performance for a synthetic metagenome with artificial in-frame stop codons

Organism	$Sens_{all}$	$Sens_{nh}$	Spec	SC predicted
Bacteria				
Alphaproteobacteria				
C. pelagibacter	88.87	92.20	97.58	71.71
R. sphaeroides	61.62	77.18	97.33	73.10
Gammaproteobacteria				
S. oneidensis	83.90	94.24	89.58	82.15
T. crunogena	64.95	79.01	97.88	80.23
V. cholerae	68.89	86.94	94.00	79.52
Cyanobacteria				
P. marinus	88.97	90.18	88.51	74.71
Synechococcus sp.	78.74	92.69	85.65	70.75
Archaea				
Euryarchaeota				
P. horikoshii	29.27	65.20	98.69	80.97
Crenarchaeota				
S. solfataricus	25.96	42.71	99.02	81.41
Average	65.69	80.04	94.25	77.17

$Sens_{all}$ is the sensitivity calculated over all genes contained in the synthetic metagenome. $Sens_{nh}$ is the sensitivity calculated over the subset of all non-hypothetical genes. SC predicted: percentage of true positives (TP) that correctly span in-frame stop codons.

Table 4. Performance for a synthetic metagenome with artificial frame shifts

Organism	$Sens_{all}$	$Sens_{nh}$	Spec	Percentage of TP predictions correctly identified as frame shift
Bacteria				
Alphaproteobacteria				
C. pelagibacter	86.39	89.32	97.73	57.71
R. sphaeroides	56.80	72.08	98.03	92.22
Gammaproteobacteria				
S. oneidensis	81.15	90.93	93.02	68.65
T. crunogena	59.69	73.33	98.33	66.67
V. cholerae	71.54	83.05	96.19	69.11
Cyanobacteria				
P. marinus	84.65	88.88	92.62	58.77
Synechococcus sp.	72.46	90.39	91.02	79.19
Archaea				
Euryarchaeota				
P. horikoshii	26.09	54.42	96.45	54.64
Crenarchaeota				
S. solfataricus	23.23	39.41	98.80	48.51
Average	62.44	75.76	95.80	66.16

$Sens_{all}$ is the sensitivity calculated over all genes contained in the contigs. $Sens_{n\ h}$ is the sensitivity calculated over the subset of non-hypothetical genes

added to each of the genes of the synthetic metagenome. For this data set, those predictions that do not match a fragment of an annotated gene where counted as false positives (FP). For those annotated genes of which at least one of its fragments is identified were counted as true positives (TP). Compared to the synthetic metagenome with no artificial mutations, the sensitivity and specificity is again only slightly reduced (Table 4). For this data set, 66% of the identified genes (TP) were also correctly predicted to have a frame shift. Noteworthy, for the synthetic metagenome without artificial errors only 357 frame shifts out of 11,686 true positive predictions were registered. This finding shows the high reliability of the method to predict frame shifts. As for the above experiments, the specificity values obtained by each genome are high, the average specificity value is 95%.

Gene identification in environmental samples obtained by 454 technology

At present, the main drawback of the recently developed high throughput parallel pyrosequencing is the short length of the reads obtained (≈100 bp on average). This is particularly undesirable when dealing with environmental data sets, since the sample is a large mixture of different species. To verify whether our algorithm is still able to identify genes in metagenomes obtained with the 454 technology, we assembled the 454 reads from the Solar Salterns sample into contigs and predicted the genes for the subset of all *long contigs*. For this verification we performed two experiments: First, a BLAST search against a database made from the set of *all contigs* from the Solar Salterns sample was conducted. Second, a direct BLAST search against a database of all 454 reads without prior assembly was employed. To validate the outcome from both experiments the respective predictions (extended to the longest possible ORF for that prediction) were compared with known proteins from the KEGG database (Ogata *et al*. (1999)) using BLAST.

Table 5. KEGG supported predictions. Number of predicted genes for a metagenome sequenced with 454 technology that have hit in the KEGG database.

Database	Number of predictions	Number of predictions with E-value up to 10^{-50}	10^{-20}	10^{-10}	10^{-5}
KEGG					
Contigs	3219	467	1544	2451	2858
Reads	3496	556	1699	2585	3044

Assembled contigs and 454 reads without prior assembly were used for BLAST search.

For both experiments, a high fraction of the predicted genes has significant BLAST hits against known proteins from the KEGG database. Remarkably, the number of predicted genes for the reads without assembly does not differ much when compared to the contigs (see Table 5). It should be pointed out that when looking at the BLAST hits against the KEGG database it seems that many of the predicted genes are fragmented due to internal frame shifts. Therefore during the BLAST search against the KEGG database, weaker E-values are obtained for these fragments. The predicted genes that do not match any known protein in the KEGG database constitute an interesting set for further studies as they could be either of false predictions, known genes with no or only a weak sequence similarity to the genes contained in KEGG, or more interestingly novel unknown genes. These results for the Solar Salterns sample demonstrate that the novel algorithm is well

Table 6. Performance for a synthetic metagenome evaluated on sequences obtained by pyrosequencing

Organism	$Sens_r$	$Sens_c$	$Sens_{nhr}$	$Sens_{nhc}$	$Spec_r$	$Spec_c$
Bacteria						
Alphaproteobacteria						
C. pelagibacter	38.96	28.02	43.92	31.66	85.29	91.54
R. sphaeroides	27.74	19.13	38.03	26.81	70.67	87.18
Gammaproteobacteria						
S. oneidensis	25.37	16.19	38.23	25.12	80.86	88.21
T. crunogena	36.21	23.49	46.31	30.29	86.42	93.43
V. cholerae	29.71	18.84	43.19	28.69	82.22	90.70
Cyanobacteria						
P. marinus	30.40	20.94	43.08	29.91	89.67	91.53
Synechococcus sp.	23.24	17.01	43.67	31.82	77.14	87.73
Archaea						
Euryarchaeota						
P. horikoshii	33.61	31.87	66.80	62.60	89.64	94.67
Crenarchaeota						
S. solfataricus	26.35	25.15	41.97	41.32	90.34	95.46
Average	30.18	22.29	45.02	34.25	83.58	91.16

$Sens_r$ and $Sens_c$ is the sensitivity for the synthetic metagenome when blasting against all 454 reads or against all assembled contigs. $Sens_{nhr}$ and $Sens_{nhc}$ is the sensitivity calculated for the subset of non-hypothetical genes of the synthetic metagenome when a BLAST search is done against the 454 reads and assembled contigs, respectively.

suited to predict genes in 'real' metagenomes, even if these samples are sequenced using the 454 technology.

Gene prediction in synthetic metagenomes using contigs and reads derived by pyrosequencing

We further evaluated the performance of the new gene finding algorithm for sequences obtained with the 454 technology (see Table 6), taking the synthetic metagenome dataset as a controlled standard of truth. The genes were predicted for the synthetic metagenomes dataset by employing a BLAST search against two different databases: one containing all assembled contigs from the Solar Salterns sample, and another containing all unassembled reads from the same sample.

In respect to the small size of the database used in the BLAST search (16 Mb for the assembled contigs and 60 Mb for the reads without prior assembly) the sensitivity obtained is very good. The highest sensitivity is reached for *Pyrococcus horikoshii*, 67% and 63% (for the subset of all non-hypothetical genes) calculated for the reads without assembly and the assembled contigs, respectively. Interestingly, these findings indicate that in contrast to the sample from the Sargasso Sea, the Archaea group is more abundant in the sample from the Solar Salterns. A second interesting observation is the good performance when running BLAST against the 454 reads without assembly, despite the fact that the average length of the reads is 100 bp. A specificity of 84% is achieved on average. Moreover, when compared to the assembled contigs the sensitivity is increased by ≈11%. In particular, these results for the short 454 reads reveal one of the strengths of our method: to consider all BLAST hits at the same time by calculating the optimal path through the matrix of combined scores instead of analyzing simple pairwise BLAST hits. This strategy allows us to identify genes that get only several short hits, even if all of the single hits are not significant.

Yet, determining the correct boundaries of the CDS when running BLAST against a small database of 454 reads is difficult, many genes are only partially covered by hits. As an optional postprocessing step our algorithm therefore can automatically extend predictions to the longest possible ORF.

Time efficiency of the novel algorithm

The running time of the novel algorithm highly depends on the size of the BLAST database since most of the running time is consumed during the BLAST based search for conserved regions, for the parsing of BLAST results as well as for the calculation of combined scores. For the evaluation presented in this survey all runs of the algorithm were executed on a compute cluster located at the Center of Biotechnology (CeBiTec), Bielefeld University. The cluster is composed of 128 Sun Fire V20z nodes. Each node has two 1.8 GHz AMD Opteron 244 CPUs and 2 Gb of RAM. The overall running time was 1 hour and 50 minutes for predicting the genes of the synthetic metagenome (≈24 Mb) when a BLAST search against half of the Sargasso Sea sample (≈390 Mb) was employed. The running time in average is 28s for the BLAST search, 17s for parsing the BLAST results and calculating the combined scores and 1s for predicting coding sequence by dynamic programming and postprocessing for a 4 Kb fragment when run on a single node using one CPU.

GLIMMER performance on synthetic metagenome

Most of the contemporary gene finding methods model frequencies of short oligonucleotides to discriminate between coding and non-coding sequences (e.g. by using a Markov chain or a Hidden Markov model). Before these methods can be used for gene prediction, usually as a first step the model needs to be trained to learn the organism specific sequence composition of the genome under study. As most of these methods model average sequence properties they may fail to adequately learn the oligonucleotide frequencies of diverse microbial assemblages. Pitfalls of existing gene finding technologies were examined by employing the state-of-the art microbial gene finder GLIMMER as an example. GLIMMER was trained on the synthetic metagenome itself as described in the Methods section. Subsequently, the trained GLIMMER was applied on each fragment. Although GLIMMER is very accurate for complete genomes (http://www.cbcb.umd.edu/software/glimmer/) the accuracy for the synthetic metagenome is strongly reduced (Table 7). Table 7 also points to one substantial problem that may affect intrinsic methods when applied to environmental data: the diverse compositional biases of different organisms contained in the sample. Another problem may be the unequal abundance of species, as overrepresented species have a stronger influence during training which may result in an unbalanced model. Also the synthetic metagenome on which GLIMMER was trained is unbalanced as it contains fragments from seven genomes with a low GC content and from two genomes with a high GC content (GC > 55%). The average GC content is ≈47%. The prediction accuracy of GLIMMER for the synthetic metagenome strongly depends on whether the fragments come from a genome with a high or low GC content. While GLIMMER has a good performance for the genomes with a low GC content, for the two genomes with a high GC content the performance is highly reduced. For the

Table 7. GLIMMER performance for a synthetic metagenome

Organism	Contig size (Mb)	GC (%)	$Sens_{all}$	$Sens_{nh}$	Spec
C. pelagibacter	1.3	30	94.34	95.54	78.39
P. marinus	1.7	31	91.80	94.87	74.04
S. solfataricus	3.0	36	91.46	93.89	72.29
P. horikoshii	1.7	42	88.28	95.60	76.92
T. crunogena	2.4	43	98.72	99.12	71.10
S. oneidensis	5.0	46	95.61	98.08	67.22
V. cholerae	3.0	48	90.68	98.48	69.52
Synechococcus sp.	2.4	59	43.80	51.60	60.41
R. sphaeroides	3.2	69	12.16	14.65	23.69
Average	2.6	47	78.53	82.42	65.95

Genomes ordered by GC content. GLIMMER was trained on the synthetic metagenome itself. $Sens_{all}$ and $Sens_{n\ h}$ is the GLIMMER sensitivity for set of all and for the subset of non-hypothetical genes. Spec: Specificity

genome with the highest GC content (*R. sphaeroides*) the accuracy is close to the one expected by a random decision drawn by a flipping a coin experiment. Owing to the diverse composition, high species richness and unequal species abundance, real metagenomes isolated from natural occurring organism assemblages possess a considerably higher complexity than the synthetic metagenome used in this study. Therefore, it is reasonable to expect that for real metagenomes the problems that affect intrinsic methods should be even more profound.

5 DISCUSSION

In this paper we presented a novel algorithm that was designed to predict genes in environmental samples. The algorithm is robust for the most common problems encountered when predicting genes in these data sets: short length of the assembled contigs and a low sequence quality.

Although, the focus of the algorithm is directed on the detection of novel genes, our algorithm can also be used to identify known genes in environmental samples: instead of searching against a database containing all fragments from the environmental sample a direct search against a database containing the sequences of known genes can be conducted.

Our results show that for large samples like the Sargasso Sea, a high fraction of the gene content can be identified based on the search for sequence conservation within the sample.

The results further demonstrate that even the short reads obtained by pyrosequencing can be used to identify protein coding genes. Therefore, environmental samples sequenced with the 454 technology may be a valuable resource to identify unknown (habitat-specific) genes. To search for novel genes our algorithm requires that at least a fraction of reads is assembled into contigs. Subsequently the complete database of reads can be used to predict the genes of these contigs. Our results therefore suggest the following strategy to identify novel (habitat-specific) genes in environmental samples: to sequence part of the sample with conventional methods to obtain longer fragments that can be assembled into contigs and additionally to sequence large amounts of data at low cost with the 454 technology to increase the size of the database that can be used to search for conserved sequences.

As our method relies on sequence similarities for the prediction of protein coding genes when running BLAST against the sample itself, the method strongly depends on the size and species composition of the sample. The sensitivity of the algorithm may be improved by incorporating general sequence properties of coding sequences or proteins.

ACKNOWLEDGEMENTS

LK was supported by the DFG Graduiertenkolleg 635 Bioinformatik. RAE and FR were supported by a grant NSF DEB-BE 04-21955 from the NSF Biocomplexity program. We thank Beltran Rodriguez-Brito for generating the environmental data. NND was supported by the Deutscher Akademischer Austausch Dienst. Thanks to the anonymous reviewers for valuable comments and helpful remarks.

REFERENCES

Badger,H. and Olsen,G.J. (1999) Article title. *Mol. Biol. Evol*, **16**, 512–524.

Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, **27**, 3911–3920.

Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. *Genome Res.*, **8**, 175–185.

Breitbart,M., Salamon,P., Andresen,B., Mahaffy,J.M., Segall,A.M., Mead,D., Azam,F. and Rohwer,F. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*, **99**, 14250–14255.

Chothia,C., Gough,J., Vogel,C. and Teichmann,S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.

Delcher,A.L., Harmon,D. and Kasif,S. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat Rev Microbiol*, **3**, 504–510.

Edwards,R.A., Rodriguez-Brito,B., Wegley,L., Haynes,M., Breitbart,M., Peterson,D., Saar,M., Alexander,S., Alexander,E.C. and Rohwer,F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics*, **7**, 57.

Frishman,D., Mironov,A., Mewes,H. and Gelfand,M. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **26**, 2941–2947.

Furrie,E. (2006) A molecular revolution in the study of intestinal microflora. *Gut*, **55**, 141–143.

Gouy,M. and Gautier (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–7074.

Green,P. (1994) Documentation for PHRAP. http://www.genome.washington.edu/UWGC/analysistools/phrap.htm.

Lombardot,T., Kottman,R., Pfeffer,H., Richter,M., Teeling,H, Quast,C. and Gloeckner,F.O. (2006) Mex.net-database resources for marine ecological genomics. *Nucleic Acids Res.*, **34**, D390–D393.

Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z., Dewell,S.B., Du,L., Fierro,J.M., Gomes,X.V., Godwin,B.C., He,W., Helgesen,S., Ho,C.H., Irzyk,G.P., Jando,S. C., Alenquer,M.L., Jarvie,T.P., Jirage,K.B., Kim,J.B., Knight,J.R., Lanza,J.R., Leamon,J.H., Lefkowitz,S.M., Lei,M., Li,J., Lohman,K.L., Lu,H., Makhijani,V.B., McDade,K.E., McKenna,M.P., Myers,E.W., Nickerson,E., Nobile,J.R., Plant,R., Puc,B.P., Ronan,M.T., Roth,G.T., Sarkis,G.J., Simons,J.F., Simpson,J.W., Srinivasan,M., Tartaro,K.R., Tomasz,A., Vogt,K.A., Volkmer,G.A., Wang,S.H., Wang,Y., Weiner,M.P., Yu,P., Begley,R.F. and Rothberg,J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Moore,J. E. and Lake,J. A. (2003) Gene structure prediction in syntenic DNA segments. *Nucleic Acids Res.*, **31**, 7271–7279.

Nekrutenko,A., Chung,W.Y. and Li,W.Y. (2003) An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet*, **19**, 306–310.

Nekrutenko,A., Chung,W.Y. and Li,W.Y. (2003) ETOPE: evolutionary test of predicted exons. *Nucleic Acids Res.*, **31**, 3564–3567.

Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **27**, 29–34.

Pruitt,K., Tatusova,T. and Maglott,R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, 501–504.

Schloss,P.D. and Handelsman,J. (2003) Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol*., **14**, 303–310.

Tringe,S. G. and Rubin,E. M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet*., **6**, 805–814.

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

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

Paper V

Releasing Metagenomics Data

Whole community genome sequencing is revolutionizing microbial ecology and fueling the next microbial genomics revolution. New sequencing technologies have made it easy for individual laboratories to produce giga-basepair (Gbp) data sets. Two of the limiting resources in handling environmental sequence data are the computer analyses and understanding of the vagaries of sequence qualities. We propose a general outline for metagenome processing and have released all of our metagenome data sets to stimulate scientific inquiry.

The first step in any analysis should be guaranteeing the congruency of the data by removing duplicate sequences, which are intrinsic sequencing artifacts, and providing unique sequence identifiers. Comparisons against the nonredundant protein database and boutique databases (e.g., rDNA, phages, mitochondria, and chloroplast databases) provide functional annotations (*1*). In our implementation, initial screens are performed with standard BLAST algorithms while more precise comparisons are made with HMM searches. The results are provided from the SEED database, an open source, freely available genome analysis database (*2*). Finally, assembling sequences into larger contiguous fragments (contigs) provides estimates of the diversity of organisms in the different environments (*3*).

The SDSU Center for Universal Microbial Sequencing Web site (SCUMS; http://scums.sdsu.edu) is an implementation of this approach. We have integrated analyses of more than 100 environmental sequence data sets totaling almost 2.5 Gbp of DNA sequence. The sequences are a combination of in-house samples, previously published data, and contributions from the metagenomics community at large, generated by either pyrosequencing or Sanger sequencing. Several mathematical tools are available on the SCUMS system, allowing rapid modeling of the communities, estimation of relative abundance and richness of organisms in the environments (PHACCS), and statistical analysis of the presence of subsystems in the samples (XIPE).

These data have the potential to shed light on natural communities with relevance to human health, global warming, coral disease, marine ecosystems, and many other areas of microbiological research. This microbial genomics revolution requires interdisciplinary help from all realms of science, particularly engineering, computer science, and statistics. By releasing all of our data, we hope interactions between scientists will be promoted, allowing the great biological questions of our age to be addressed.

Forest Rohwer,[1,2]* Hong Liu,[1] Florent E Angly,[1,3] Steven Rayhawk,[1,3] Lutz Krause,[4] Robert Olson,[5] Beltran Rodriguez Brito,[1,3] Rick Stevens,[5] Robert A. Edwards[1,2,3,6]*

[1]Department of Biology, San Diego State University, San Diego, CA, USA. [2]Center for Microbial Sciences, San Diego State University, San Diego, CA, USA. [3]Computational Science Research Center, San Diego State University, San Diego, CA, USA. [4]Bielefeld University, Center for Biotechnology (CeBiTec), D-33594 Bielefeld, Germany. [5]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA. [6]Fellowship for Interpretation of Genomes, Burr Ridge, IL, USA.

*To whom correspondence should be addressed. Email: frohwer@gmail.com (F.R.); raedwards@gmail.com (R.E.)

References

1. See http://phage.sdsu.edu/phage.
2. See http://www.theseed.org/; http://metagenomics.theseed.org/.
3. F. E. Angly et al., PLoS Biol. 4, e368 (2006).

Paper VI

Taxonomic Classification of Short Environmental DNA Fragments

Lutz Krause [a], Naryttza N. Diaz [a], Alexander Goesmann [a], Forest Rohwer [b,c], Scott Kelley [c], Robert A. Edwards [b,c,d], Jens Stoye [a,e]

May 31, 2007

[a]Bielefeld University, Center for Biotechnology (CeBiTec), D-33594 Bielefeld, Germany
[b]Center for Microbial Sciences, San Diego, CA
[c]Department of Biology, San Diego State University, San Diego, CA
[d]Mathematics & Computer Science Division, Argonne National Laboratory, Argonne, IL
[e]Universität Bielefeld, Technische Fakultät, D-33594 Bielefeld, Germany

Abstract

Metagenomics is providing striking insights into the ecology of microbial communities. The recently developed massively parallel 454 pyrosequencing technique gives the opportunity to rapidly obtain metagenomic sequences at a low cost and without cloning bias. However, the taxonomic analysis of the short reads produced represents a significant computational challenge. A phylogenetic algorithm for predicting the source organisms of environmental 454-reads is described. The algorithm searches for conserved Pfam domain and protein families in the unassembled reads of a sample. These gene fragments (*environmental gene tags, EGTs*), are classified into taxonomic groups based on the reconstruction of a phylogenetic tree of each matching

Pfam family. The method exhibits a high accuracy for a wide range of taxonomic groups and EGTs as short as 33 amino acids can be taxonomically classified up to the rank of genus. The phylogenetic algorithm was applied in a comparative study of three aquatic microbial samples that were obtained by 454 pyrosequencing. Profound differences in their species composition could clearly be revealed.

Introduction

In metagenomics, the collective genomes from natural microbial communities are randomly sampled from the environment and subsequently sequenced [1, 5, 27, 29, 30]. By directly accessing the genomic DNA of coexisting microbial species, these approaches have the potential of giving a comprehensive view of the evolution, lifestyle, and diversity of free-living microbes (e.g. [6, 11, 12, 13, 19, 32]). Moreover, considering that the vast majority of microbes resists cultivation with conventional methods [14, 15, 23], metagenomics highly enlarges our window into the hidden world of microbes.

The massively parallel pyrosequencing system was recently developed by 454 Life Sciences, and has dramatically dropped the time and cost constraints of DNA sequencing [18]. Pyrosequencing not only produces large amounts of data at a low cost, but also allows sequencing of environmental DNA without a prior cloning step [8, 28]. Despite these advantages, the main drawback of the 454 technology is that at present only short reads of $\approx$100 bp are obtained. Short read length, inherent genetic heterogeneity within populations, inter-species gene conservation, and variable species richness and evenness all make the assembly of environmental 454-reads into longer contiguous DNA sequences (contigs) a fundamental computational challenge.

Assessing the taxonomic composition of microbial communities is an essential question in metagenomics; but is still in its infancy. In this study, a novel method for the taxonomic classification of un-assembled 454-reads of an environmental sample is presented.

To infer the taxonomic origin of an environmental DNA fragment, one type of approach relies on the identification of phylogenetic 'marker genes' such as rDNA or *recA* genes [30, 33].

While these methods frequently yield a high accuracy, only a small fraction of fragments can be taxonomically characterized, depending on the size of the used marker gene database. To overcome this limitation, novel methods have recently been devised that analyze the presence of short oligonucleotides or motifs to classify environmental DNA sequences into taxonomic groups [20, 26]. These methods give the capacity to accurately infer the source organisms of longer stretches of DNA, but to our knowledge cannot be applied to sequences shorter than 1,000 bp. On the other hand, simply classifying genomic fragments based on a best BLAST hit will only yield reliable results if close relatives are available for comparison [16].

The phylogenetic algorithm presented herein, uses all Pfam [10] domain and protein families as phylogenetic markers to identify the source organisms of environmental DNA fragments as short as 100 bp. The method has two components: The first, is used for identifying domain and protein family fragments in the un-assembled reads of a sample using Pfam profile hidden Markov models (pHMMs). Profile HMMs are very accurate for the detection of weak functional signals and short conserved functional sequences, which makes this technique particularly adequate for the analysis of un-assembled 454-reads. In this study, environmental domain and protein family fragments identified in a metagenome are defined as *environmental gene tags* (*EGTs*). In the second component, a phylogenetic tree is reconstructed for each matching Pfam family. Environmental gene tags are taxonomically classified based on their phylogenetic relationships to family members with known taxonomic affiliations.

The algorithm was extensively evaluated on synthetic data sets. For taxonomic groups that are well represented in the Pfam database, EGTs as short as 33 amino acids can accurately be classified with an average specificity ranging from 97% for superkingdom to 68% for genus. The average sensitivity ranges from 90% for superkingdom to 40% for genus. Moreover, the power of the method for studying the taxonomic composition of environmental samples was demonstrated in a comparative analysis of three aquatic microbial ecosystems. The analysis clearly revealed profound differences in the taxonomic composition of microbial communities from different aquatic habitats. All source code is freely available upon request to the corresponding author.

Materials and Methods

Data Sets

The Pfam fragment pHMM library (Pfam_fs), the Pfam MySQL database, the full multiple alignment of each Pfam family, as well as a fasta version of Pfam's underlying sequence database (pfamseq) were downloaded from the Pfam web site (Pfam version 20.0). Pfam families with less than 10 members were excluded from the data set. Duplicate sequences were removed from each multiple alignment: If multiple copies of the same sequence of one organism were present (e.g. 100% identical sequences from different strains), only one of these was retained.

For constructing a synthetic metagenome as a standard of truth for the performance evaluation, 77 complete genomes were downloaded from GenBank [4] and split into non-overlapping fragments of 100 bp. Genomes were included that stem from taxonomic groups that are both over- and under-represented in the Pfam database. The taxa information of the organisms was obtained from the US National Center for Biotechnology Information (NCBI) Taxonomy database [31].

The 454-reads of three 'real' microbial samples – a coral reef sample, a solar saltern sample, and a stromatolite sample – were downloaded from the SDSU Center for Universal Microbial Sequencing (SCUMS) [24]. The coral reef sample was isolated from coral reef waters at the Kingman atoll located in the northern Line Islands of the central Pacific (coordinates: -162.3347833 W 6.38566667 N; Dinsdale, *et al.*, Submitted). The solar saltern sample was collected from the solar salterns in San Diego, CA (coordinates: -117.107356 W, 32.599040 N; Rodriguez-Brito *et al.*, Unpublished). The stromatolite sample was taken from Rios Mesquites, Mexico (coordinates: -102.066390 W 26.985876 N; Desnues *et al.*, Unpublished). Total community DNA of all three samples was purified as described elsewhere [8] and in their papers, and sequenced using pyrosequencing by 454 Life Sciences, Branford, CT.

Algorithm

The presented method relies on two algorithmic components: The first is used for the detection of environmental Pfam domain and protein family fragments (EGTs) that are conserved in an environmental sample. The second reconstructs a phylogenetic tree (family tree) for each matching family. These trees consist of all previously detected EGTs matching the family (*matching* EGTs) as well as all family members with a known taxonomic origin, called *taxaknown* members. Environmental gene tags are taxonomically classified based on their location in respect to the *taxaknown* members in the reconstructed trees.

Detecting Environmental Gene Tags

Environmental gene tags are identified using the profile hidden Markov models (pHMMs) from the protein families database (Pfam). Pfam is a comprehensive database of manually curated domain and protein families [10]. Each family is represented by a full multiple alignment of all known family members as well as by a pHMM, which can be employed to search for new, unknown family members.

First, a similarity search of each read of a sample is conducted against Pfam's underlying sequence database using BLASTX [3] with the '-w' frame-shift option. This computes the 6-frame translations, predicts frame-shifts, and identifies candidate members of Pfam families. Reads without a BLAST hit of E-value ≤ 1 are excluded from further analysis. This preprocessing step highly reduces the amount of computational effort that needs to be done with the pHMMs.

After all predicted frame-shifts are resolved, remaining reads are screened for conserved Pfam domain and protein families using the highly accurate Pfam pHMMs. Reads with a BLASTX hit are aligned to the matching families using their local pHMM from the Pfam_fs database (E-value cut-off of 0.01). By using local pHMMs, even domain and protein families that are only partly covered by a read are identified. The sequences of all identified EGTs are added to the multiple alignment of the matching Pfam family using hmmalign from the hmmer package [7].

Taxonomic Classification of Short EGTs

Environmental gene tags are classified into taxonomic categories based on the reconstruction of a phylogenetic tree. The multiple alignments of *taxaknown* members and matching EGTs of each Pfam family are used to calculate a pairwise distance of all combinations of *taxaknown* members and matching EGTs. The distance between two sequences is defined as their pairwise sequence identity, i.e. the fraction of identical amino acids in the aligned region. In case that the sequences of two EGTs do not have a sufficient overlap, their distance is estimated as described in 'Estimating Distance of Non-Overlapping EGTs' below. An unrooted phylogenetic tree is reconstruct from the pairwise distances using the neighbor-joining clustering method (with the NEIGHBOR program from the PHYLIP package [9]). An adapted version of the algorithm developed by Nguyen *et al.* [21] is employed for parsing the reconstructed trees. Environmental gene tags are classified depending on their phylogenetic relationships to *taxaknown members*. If an EGT g is localized within a group of *taxaknown* members sharing a common taxon t, then g is classified as t. Otherwise, it is classified as '*unknown taxon*' (Figure 1).

In detail, let T be an unrooted, binary family tree with nodes V. For an EGT $g \in V$, let $c^*(g)$ denote the subtree of T that has the smallest number of *taxaknown* members, while at the same time fulfilling the two conditions:

1. $g \in c^*(g)$
2. $c^*(g)$ has at least three *taxaknown* members

Notably, for unrooted trees, three different subtrees arise from each internal node. For each taxonomic rank (superkingdom, phylum, class, order, and genus) if at least 80% of the *taxaknown* members of $c^*(g)$ share a common taxon t, then g is also classified as t, otherwise it is classified as '*unknown taxon*' (Figure 1). The values used for the internal parameters were determined during an optimization phase of the algorithm.

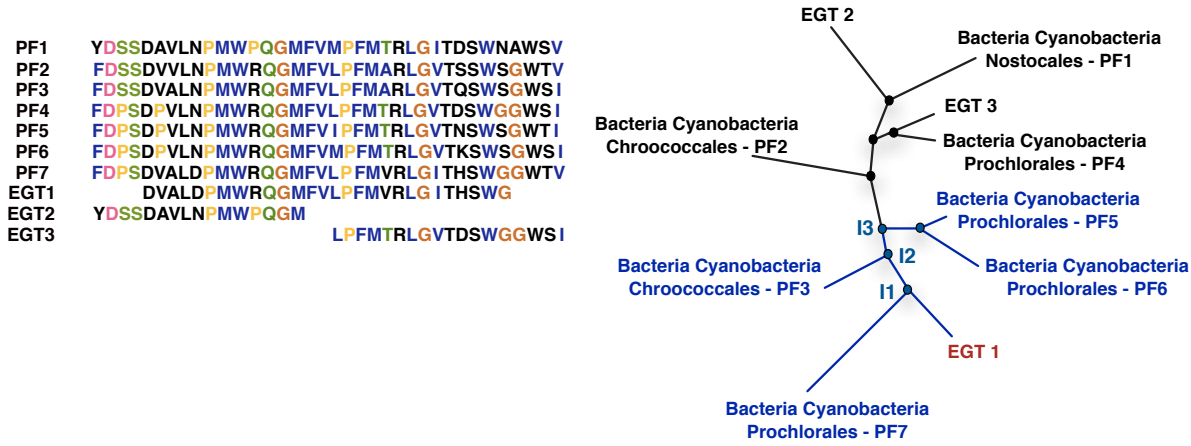


Figure 1: Unrooted phylogenetic tree reconstructed from a toy example multiple alignment. The multiple alignment shown was constructed from *taxaknown* members of a Pfam family ($PF1,\ldots,PF7$) and EGTs matching that family ($EGT1, EGT2, EGT3$). A phylogenetic tree reconstructed from the alignment is illustrated on the right. The environmental gene tag $EGT1$ is localized in a subtree $c^*(EGT1)$ of cyanobacteria (depicted in blue). Hence, it is classified as 'Bacteria Cyanobacteria'. As $c^*(EGT1)$ contains cyanobacteria from different genera $EGT1$ is classified as *unknown taxon* at the rank of genus.

Estimating Distance of Non-Overlapping EGTs

Extremely short reads (such as reads of length 100 bp) frequently only partially cover a Pfam family. Sequences of such EGTs may not overlap in the constructed multiple alignment (for example $EGT2$ and $EGT3$ in Figure 1). As the pairwise sequence identity of non-overlapping EGTs cannot be assessed from an alignment, their distance is estimated as follows: Let S be the set of all sequences contained in a multiple alignment and $d(s,s')$ be the pairwise distance of two sequences $s,s' \in S$. If the sequences $s,s' \in S$ of two EGTs overlap with less than ten amino acids, their distance is estimated by the additive estimation as proposed by Landry *et al.* [17]:

$$d(s,s') \;=\; \operatorname*{argmin}_{l,k} \; \max\{d(s,k)+d(s',l);d(s,l)+d(s',k)\} - d(l,k), \text{ where}$$

$$l,k \in S.$$

Measuring the Accuracy

The classification accuracy of the phylogenetic algorithm was evaluated on short DNA fragments with known taxonomic origins. By comparing the predicted with the known taxa, the sensitivity, specificity, false negative rate, false positive rate, and unknown rate were assessed. For a taxonomic class i, let P_i be the total number of EGTs from i; TP_i the number of EGTs that is correctly classified into i; FP_i the number of EGTs that is erroneously assigned to i; FN_i the number of EGTs from i that is misclassified into some class $j \neq i$; and U_i the number of EGTs from i that is classified as *unknown taxon*. Note that $P_i = TP_i + FN_i + U_i$. The *sensitivity* measures the proportion of EGTs that is correctly classified. For a taxonomic group i, it is defined as $Sn_i = \frac{TP_i}{P_i}$. The *specificity* measures the reliability of classifications and is defined as $Sp_i = \frac{TP_i}{TP_i+FP_i}$. The *false negative rate* is defined as $FNrate_i = \frac{FN_i}{P_i}$. It measures the proportion of EGTs from a taxonomic class i that is falsely assigned to any class $j \neq i$. The *unknown rate* measures the proportion of EGTs that cannot be taxonomically classified and is defined as $Urate_i = \frac{U_i}{P_i}$. The *false positive rate* is the proportion of EGTs that is falsely assigned to a class i. It is defined as $FPrate_i = \frac{FP_i}{\sum_{j \neq i} P_j}$.

Measuring the Accuracy for the Taxonomic Classification of Short EGTs

In a first experiment, the performance for the taxonomic classification of EGTs was evaluated in a Jack-knife approach for all taxa present in the Pfam database. As previously mentioned, each Pfam family is represented by a full multiple alignment of all known family members. All known members of each family were randomly partitioned into ten subsamples. In ten steps, one subsample was withdrawn from the full multiple alignment and classified using all remaining subsamples as follows: From each withdrawn sequence only 33 contiguous amino acids were randomly selected as *artificial EGTs*. Subsequently, artificial EGTs were again added to the multiple alignment of the remaining subsamples. Based on the resulting multiple alignment each artificial EGT was classified as described in section 'Taxonomic Classification of Short EGTs' above. The accuracy was separately evaluated at each taxonomic rank (superkingdom, phylum, class, order, and genus).

In general, the accuracy of the phylogenetic algorithm highly depends on the representation of taxa in the Pfam database. In the performance evaluation, the accuracy was separately evaluated for well represented ($\geq$ 4000 Pfam members) and for poorly represented taxa ($<$ 4000 Pfam members).

Measuring the Accuracy for 100 bp Fragments from 77 Complete Genomes

In a second experiment, the accuracy of the complete algorithm was evaluated on a synthetic metagenome consisting of 100 bp fragments from a wide range of taxonomic groups. The metagenome was created by splitting 77 complete bacterial and archaeal genomes into 100 bp, non-overlapping fragments. The taxonomy of the fragments was predicted using our complete classification algorithm: First, EGTs (fragments of Pfam families) were identified in the 100 bp fragments and subsequently classified. Usually, a high fraction of reads in environmental samples comes from genomes that have not been sequenced yet. To account for this, all known Pfam members belonging to the same species as any of the 77 complete genomes were omitted from the full multiple alignments. As a consequence, at the rank of genus a high fraction of EGTs could not be classified into their taxonomic group. In this experiment the performance was therefore evaluated only up to the rank of order.

Measuring the Diversity

The abundance and evenness of organisms in different communities can be characterized with Shannon's diversity index [25] (also called Shannon-Wiener index). In the context of this work, for a taxonomic rank r, it is defined as

$$H' = -\sum p_i \ln p_i,$$

where p_i is the proportion of EGTs that is classified into the i-th taxonomic group of rank r. The *species evenness* can then be defined as

$$J = \frac{H'}{\ln(H_{max})},$$

where H_{max} is the total number of taxa found at rank r.

Usually, diversity and evenness are measured at the rank of species. Nonetheless, as quantitative species information is not available for the three aquatic environmental samples, diversity and evenness of prokaryotes were measured at the rank of phylum, class, order, and genus.

Results

Accuracy for the Taxonomic Classification of Short EGTs

In the first experiment, the classification of short EGTs was extensively evaluated for a wide range of taxonomic categories, including DNA fragments from archaea, bacteria, eukaryotes, and viruses. On the whole, EGTs as short as 33 amino acids ($\approx$ 100 bp) can be accurately classified up to the rank of genus (Figures 2 and 3). For well represented classes (all four superkingdoms, 20 phyla, 27 classes, 59 orders, and 69 genera), between 97% at superkingdom and 68% at genus of predicted taxa were correct. The average sensitivity ranged from 90% (superkingdom) to 40% (genus). Between 7% (superkingdom) and 44% (genus) of EGTs could not be assigned to any taxonomic group and hence were classified as *unknown taxa*.

The accuracy depends on how well a taxonomic class is represented in the Pfam database. The taxa of EGTs from poorly represented classes frequently cannot be inferred from the phylogenetic tree, and so in this case, EGTs should be classified as *unknown taxa*. Overall, reliable predictions could also be obtained for poorly represented taxa, with an average specificity ranging from 84% to 65%. As expected, the average sensitivity considerably dropped (to 8-19%), while the unknown rate increased (to 34% at the rank of phylum and to 63% at the rank of genus).

The background noise that can be expected for each taxonomic group was measured by the false positive rate, i.e. the probability that an EGT is by chance falsely classified into that group. Also the false positive rate highly depends on the representation of taxa in Pfam (Figures 4 and 5). For example, an EGT was falsely assigned to the overrepresented proteobacteria with a probability of 2.3%. On the other hand, an EGT was misclassified into the less represented chloroflexi with a probability of 0.03%. In summary, for well represented taxonomic groups the average false positive rate ranged from 0.7% at the rank of superkingdom to 0.12% at genus. For poorly repre-

sented taxa the average false positive rate was below 0.004% at all taxonomic ranks. Noteworthy, particularly in light of advances in sequencing technology, longer fragments of length 200 bp and 400 bp resulted in a slightly improved accuracy (sensitivity, specificity, false negative rate, and unknown rate) (data not shown).

Accuracy for 100 bp Fragments from 77 Complete Genomes

In a second experiment, the complete algorithm – i.e. the detection of EGTs followed by their taxonomic classification – was evaluated on a synthetic metagenome consisting of 100 bp fragments from 77 complete genomes. These genomes covered both archaea and bacteria, 10 phyla, 11 classes, 29 orders, and 62 genera, thus representing the metagenome of a complex microbial community. In light of the short fragment length, also in this experiment a high classification accuracy was achieved (Figure 6). On average, 81% (superkingdom) to 41% (order) of identified EGTs were correctly assigned to their corresponding taxon, while 9% (superkingdom) to 16% (order) were misclassified. Between 11% and 41% of EGTs could not be taxonomically assigned and thus were classified as *unknown taxa*.

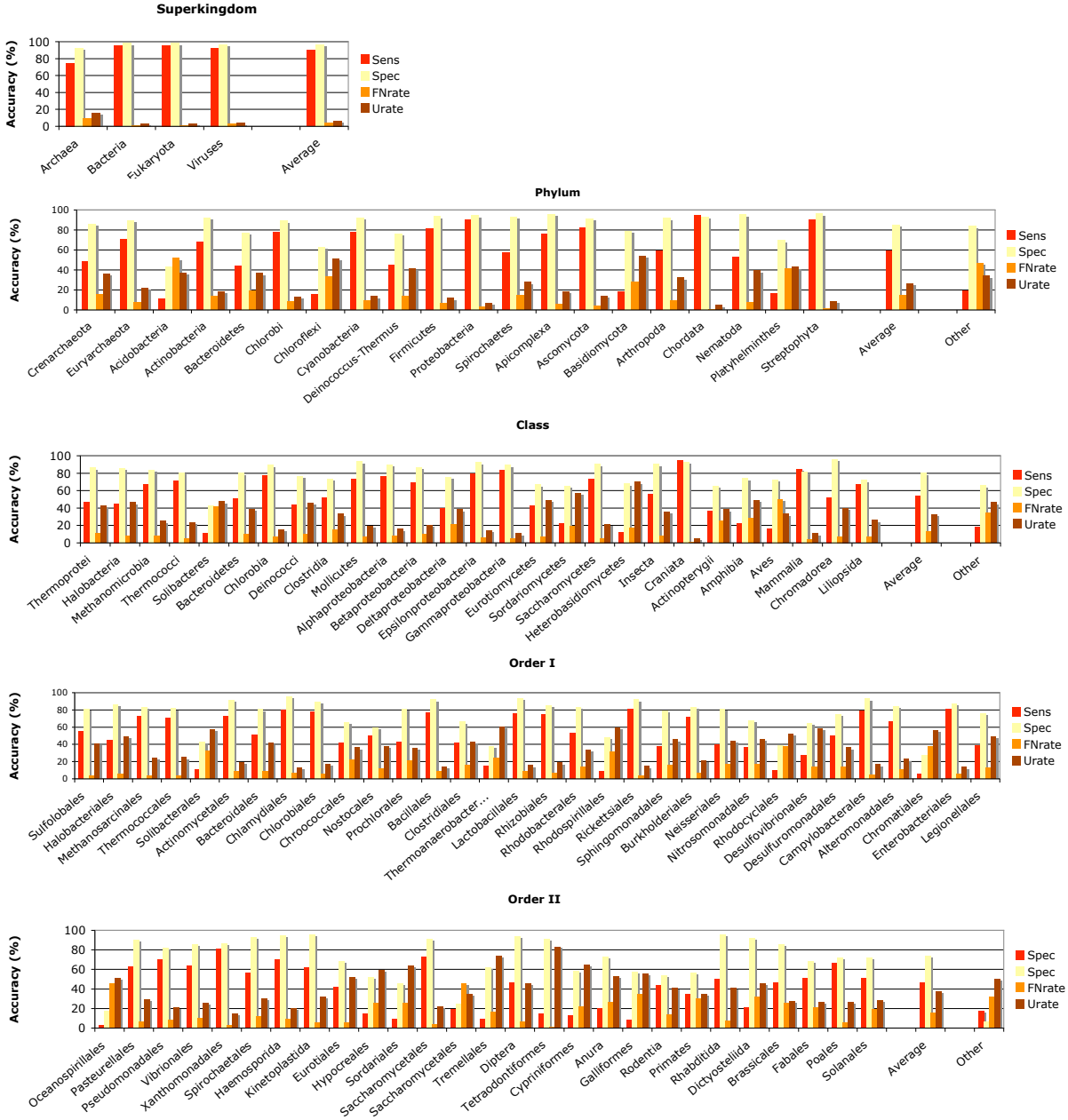


Figure 2: Rank dependent classification accuracy for 33 amino acid fragments. For each taxonomic rank from superkingdom to order the sensitivity (Sens), specificity (Spec), false negative rate (FNrate), and unknown rate (Urate) are shown. 'Other' depicts the average accuracy for taxa that are poorly represented in Pfam.

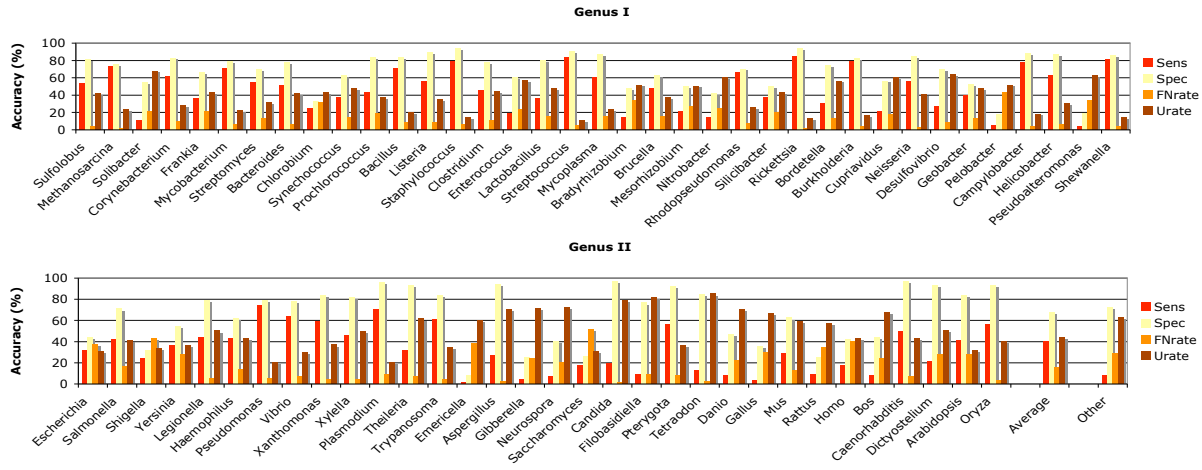


Figure 3: Classification accuracy for 33 amino acid fragments at the rank of genus. The sensitivity (Sens), specificity (Spec), false negative rate (FNrate), and unknown rate (Urate) are shown. 'Other' depicts the average accuracy for taxa that are poorly represented in Pfam.

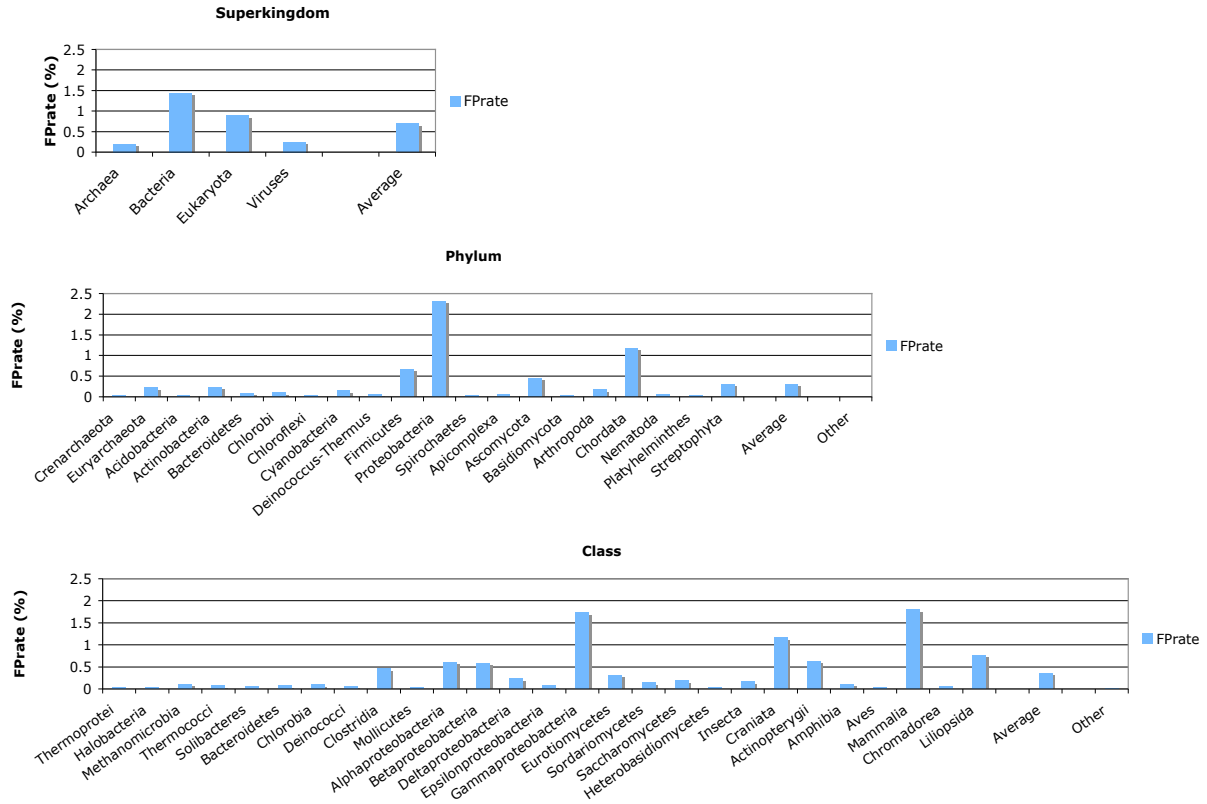


Figure 4: False positive rate for 33 amino acid fragments at the rank of superkingdom, phylum, and class. 'Other' depicts the average false positive rate for taxa that are poorly represented in Pfam.

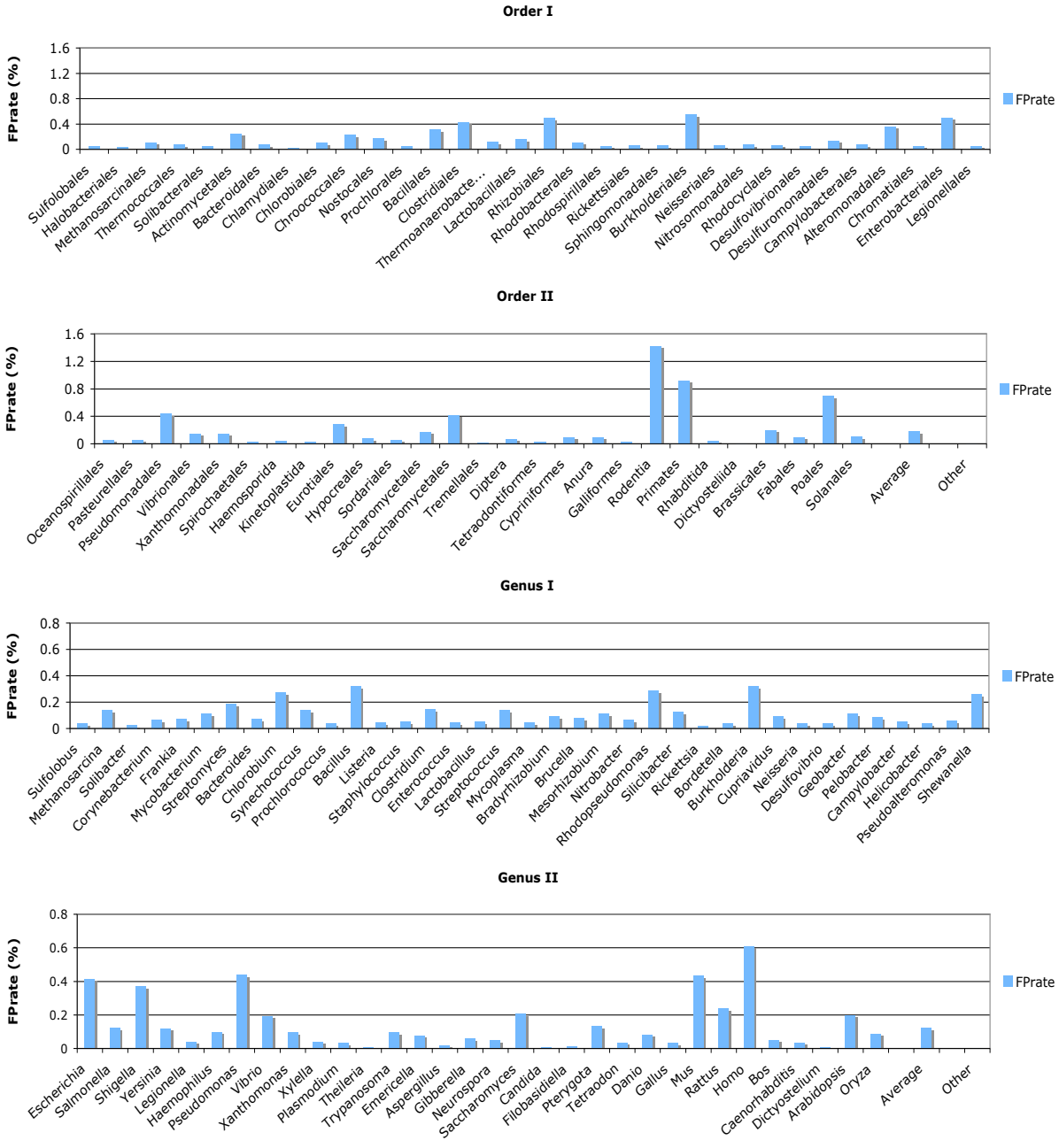


Figure 5: False positive rate for 33 amino acid fragments at rank of order and genus. Note that at order the scale ranges from 0 to 1.6, while at genus it ranges from 0 to 0.8. 'Other' depicts the average false positive rate for taxa that are poorly represented in Pfam.

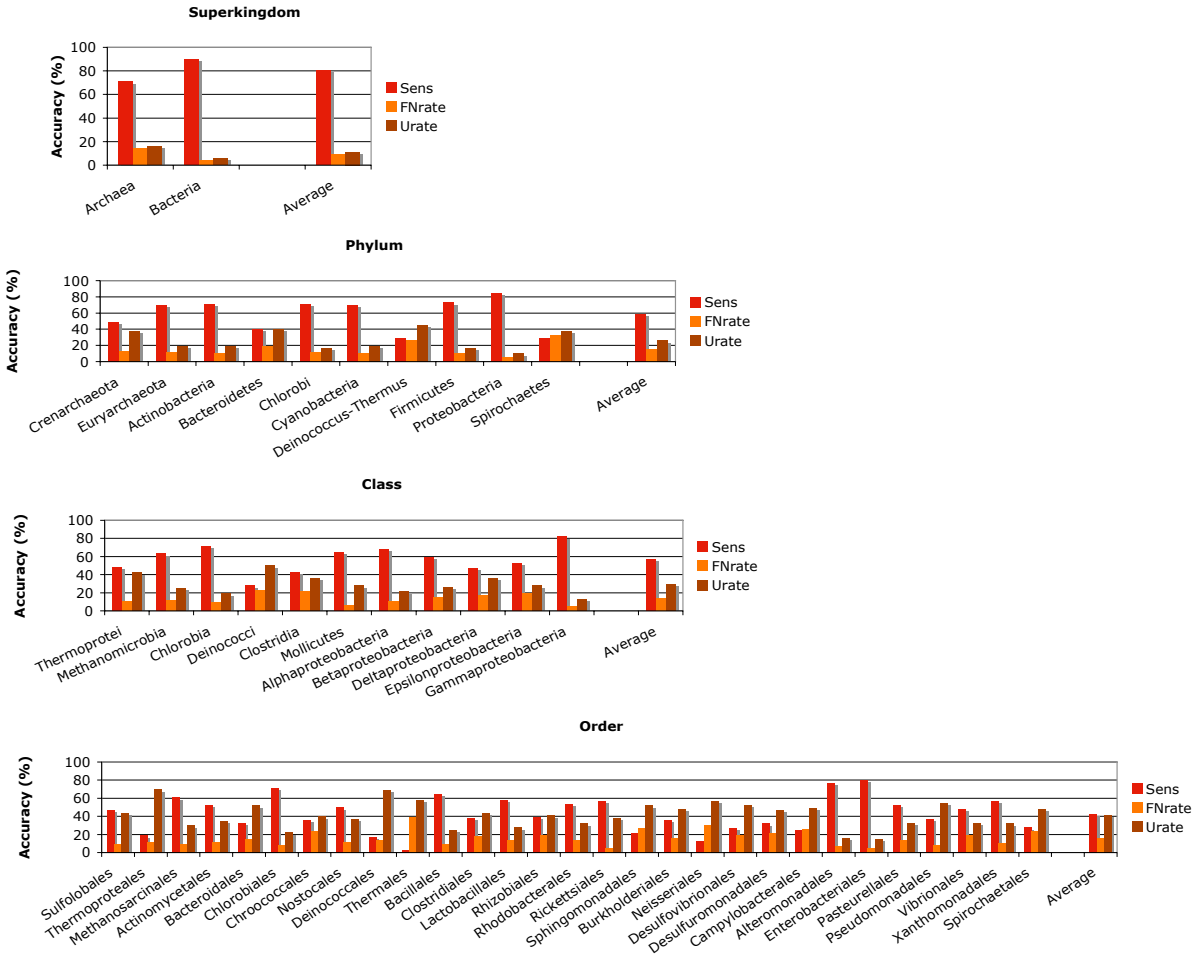


Figure 6: Accuracy of the taxonomic assignment of 100 bp fragments from 77 complete genomes. The sensitivity (Sens), false negative rate (FNrate), and proportion of EGTs that could not be assigned to any taxonomic group (Urate) are shown.

Comparative Analysis of Microbial Communities from Disparate Aquatic Environments

To identify taxonomic trends in microbial communities from disparate aquatic environments, the method presented herein was applied in a comparative analysis of three short-read metagenomes isolated from Kingman coral reefs, San Diego solar salterns, and Rios Mesquites stromatolites. All three samples were sequenced using the 454 pyrosequencing technology. Using our phylogenetic algorithm, a high proportion of EGTs was taxonomically assigned, ranging from 75-92% for superkingdom to 33-42% for genus (Table 1). The taxonomic characterization of the samples indicated a significant difference in species composition (Figure 7). In contrast to the coral reef and stromatolite samples, where bacteria dominated (68% and 79% of EGTs), 49% of EGTs from the solar saltern sample were classified as archaea and only 20% as bacteria.

For the prokaryotic fraction of EGTs (pEGTs), the coral reef sample had the highest predicted diversity and evenness (Table 2). While proteobacteria was the most abundant phylum (59% of pEGTs), a significant proportion of pEGTs was also assigned to actinobacteria (4% of pEGTs), bacteroidetes (4% of pEGTs), cyanobacteria (4% of pEGTs), firmicutes (4% of pEGTs), and planctomycetes (3% of pEGTs). At the rank of order and genus, the coral reef sample was highly diverse, with rhodobacterales (11% of pEGTs) being the most prevalent order and *Silicibacter* (5% of pEGTs) and *Pirellula* (3% of pEGTs) the most abundant genera.

The stromatolite sample had an intermediate diversity and evenness for the prokaryotic fraction of EGTs (Table 2). At the rank of phylum, it was mainly dominated by cyanobacteria (57% of pEGTs). Additionally, a considerable fraction of pEGTs was classified as proteobacteria (15% of pEGTs) and firmicutes (4% of pEGTs). Nostocales (20% of pEGTs) and chroococcales (17% of pEGTs) were the most abundant orders.

The solar saltern sample had the lowest prokaryotic diversity and evenness (Table 2). The majority of pEGTs was assigned to different halobacteria (58% of pEGTs), namely *Natronomonas* (14% of pEGTs), *Haloarcula* (12% of pEGTs), *Halobacterium* (8% of pEGTs), and *Haloferax*

(1% of pEGTs). At the rank of phylum, euryarchaeota (69% of pEGTs) was the most prevalent group followed by proteobacteria (12% of pEGTs). The remaining phyla were only poorly represented ($\leq$ 2% of pEGTs).

The results clearly revealed differences in the cyanobacterial composition between the coral reef and stromatolite environments (Figure 7). *Synechococcus*-like species were predicted to be the most prevalent cyanobacteria in the stromatolite sample (6% of pEGTs), but *Prochlorococcus*-like species were predicted to be the dominant cyanobacteria in the coral reef sample (2% of pEGTs). At the rank of genus, a considerable number of pEGTs from the stromatolite sample was assigned to diverse genus of the cyanobacteria group: *Synechococcus* (6%), *Nostoc* (5%), *Crocosphaera* (4%), *Anabaena* (4%), *Gloeobacter* (1%), *Synechocystis* (2%), *Trichodesmium* (2%), and *Prochlorococcus* (0.7%). In contrast, for the coral reef sample *Prochlorococcus* (2% of pEGTs), *Synechococcus* (0.2% of pEGTs), and *Synechocystis* (0.2% of pEGTs) were the only cyanobacteria with a considerable number of assigned pEGTs.

These findings reflected the environments where the samples were collected. Marine microbial communities have been reported as complex and diverse, with a high proportion of proteobacteria and a considerable number of cyanobacteria (*Prochlorococcus* and *Synechococcus*) [30]. The stromatolites were formed by cyanobacteria [2]. However, compared to some earlier studies of stromatolites (e.g. [22]), the proportion of cyanobacteria predicted in the Rios Mesquites stromatolite was remarkably high. On the other hand, the high proportion of different halobacteria found in the solar saltern sample reflected the stress condition caused by high salt concentration, shaping the community composition of this habitat.

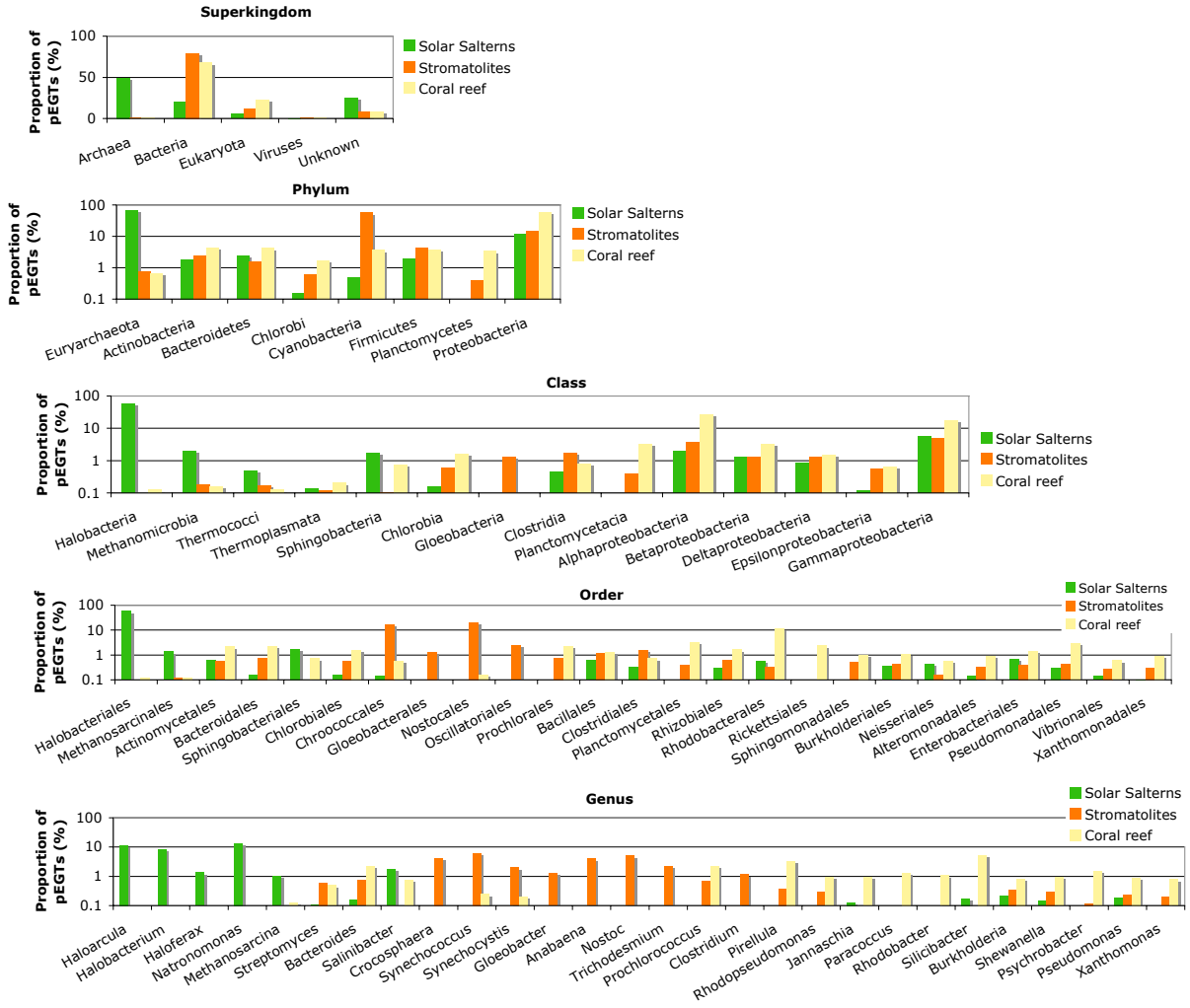


Figure 7: Taxonomic characterization of three environmental samples obtained by 454 pyrosequencing.

Sample	Size	EGTs	Proportion of EGTs taxonomically assigned				
			Superkingdom	Phylum	Class	Order	Genus
Coral reef	188.445	3.577	75%	66%	53%	53%	33%
Stromatolite	124.694	7.414	92%	77%	72%	70%	37%
Solar Saltern	582.681	55.605	92%	71%	57%	56%	42%
Average			86%	68%	61%	60%	37%

Table 1: Taxonomic characterization of three metagenomes obtained by 454 pyrosequencing. The sample size (number of reads), the number of EGTs identified in each sample, and the proportion of EGTs that were taxonomically assigned at the rank of superkingdom, phylum, class, order, and genus are shown.

Sample	Phylum		Class		Order		Genus	
	H'	J	H'	J	H'	J	H'	J
Coral reef	1.2	0.46	1.7	0.55	3.9	0.81	4.2	0.83
Stromatolite	1.1	0.42	1.16	0.37	2.7	0.55	3.6	0.70
Solar Saltern	0.8	0.31	1.0	0.32	1.4	0.28	2.6	0.45

Table 2: Prokaryotic diversity (H') and evenness (J) in three aquatic microbial samples at rank of phylum, class, order, and genus.

Conclusion

A novel method was developed for predicting the taxonomic origins of short environmental DNA fragments. In the first phase, domain and protein family fragments (environmental gene tags, EGTs) are identified in the un-assembled reads of a sample using Pfam profile hidden Markov models. In the second phase, a phylogenetic tree (family tree) is reconstructed for each matching Pfam family. Environmental gene tags are classified based on their location in the respective family tree. With this strategy, families that are not suited to infer phylogenies, such as rapidly evolving families or families with members that are frequently inherited by horizontal gene transfer, are implicitly identified. Trees reconstructed from these families have 'mixed subtrees' with members from various different taxa. In this case, the contained EGTs are classified as '*unknown taxa*'.

The results shown in this study clearly demonstrate that short fragments of Pfam domain and protein families are well suited as phylogenetic markers for inferring the taxonomic affiliations of short environmental DNA fragments. In comparison to methods that rely on only a few marker genes, such as 16S rDNA or *recA* genes, the use of all Pfam families provides a deeper picture into the taxonomic composition of microbial samples. In this work, the comparative study of three aquatic microbial communities is an example on how the predicted taxa yield detailed insights into the species composition of environmental samples obtained by 454 pyrosequencing.

Acknowledgments

LK was supported by the DFG Graduiertenkolleg 635 Bioinformatik and by the International NRW Graduate School in Bioinformatics and Genome Research. Parts of this work were conducted during a research visit of LK at Rob Edwards' and Forest Rohwer's groups, San Diego State University, CA. We thank Christelle Desnues, Elizabeth Dinsdale, and Beltran Rodriguez-Brito for sharing data prior to publication. We would also like to acknowledge Eric R. Alegre for his help developing the original tree-parsing algorithm.

References

[1] E. E. Allen and J. F. Banfield. Community genomics in microbial ecology and evolution. *Nat Rev Microbiol*, 3:489–498, 2005.

[2] A. C. Allwood, M. R. Walter, B. S. Kamber, C. P. Marshall, and I. W. Burch. Stromatolite reef from the Early Archaean era of Australia. *Nature*, 441:714–718, 2006.

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

[4] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Res*, 35:D21–D25, 2007.

[5] M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*, 99:14250–14255, 2002.

[6] O. Béjà, L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289:1902–1906, 2000.

[7] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[8] R. A. Edwards, B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. Peterson, M. Saar, S. Alexander, E. C. Alexander, and F. Rohwer. Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics*, 7:57, 2006.

[9] J. Felsenstein. Phylip: Phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[10] R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, and

A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34:D247–D251, 2006.

[11] E. Furrie. A molecular revolution in the study of intestinal microflora. *Gut*, 55:141–143, 2006.

[12] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312:1355–1359, 2006.

[13] S. K. Hansen, P. B. Rainey, J. A. J. Haagensen, and S. Molin. Evolution of species interactions in a biofilm community. *Nature*, 445:533–536, 2007.

[14] P. Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biol*, 3:REVIEWS0003, 2002.

[15] P. Hugenholtz, B. M. Goebel, and N. R. Pace. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol*, 180:4765–4774, 1998.

[16] L. B. Koski and G. B. Golding. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 52:540–542, 2001.

[17] P.-A. Landry, F.-J. Lapointe, and J. A. W. Kirsch. Estimating phylogenies from lacunose distance matrices: additive is superior to ultrametric estimation. *Mol Biol Evol*, 13:818–823, 1996.

[18] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M.

Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[19] H. G. Martín, N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon, and P. Hugenholtz. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol*, 24:1263–1269, 2006.

[20] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, 4:63–72, 2007.

[21] T. X. Nguyen, E. R. Alegre, and S. T. Kelley. Phylogenetic analysis of general bacterial porins: a phylogenomic case study. *J Mol Microbiol Biotechnol*, 11:291–301, 2006.

[22] D. Papineau, J. J. Walker, S. J. Mojzsis, and N. R. Pace. Composition and structure of microbial communities from stromatolites of Hamelin Pool in Shark Bay, Western Australia. *Appl Environ Microbiol*, 71:4822–4832, 2005.

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

[24] F. Rohwer, H. Liu, F. E. Angly, S. Rayhawk, L. Krause, R. Olson, B. Brito, R. Stevens, and R. A. Edwards. Releasing metagenomics data. *Science*, accepted.

[25] C.E. Shannon and W. Weaver. *The mathematical theory of communication*. Urbana, Univerisity of Illinois Press, Urbana, IL, 1963.

[26] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol*, 6:938–947, 2004.

[27] S. G. Tringe and E. M. Rubin. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet*, 6:805–814, 2005.

[28] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444:1027–1031, 2006.

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

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

[31] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 35:D5–12, 2007.

[32] T. G. Whitham, J. K. Bailey, J. A. Schweitzer, S. M. Shuster, R. K. Bangert, C. J. LeRoy, E. V. Lonsdorf, G. J. Allan, S. P. DiFazio, B. M. Potts, D. G. Fischer, C. A. Gehring, R. L. Lindroth, J. C. Marks, S. C. Hart, G. M. Wimp, and S. C. Wooley. A framework for community and ecosystem genetics: from genes to ecosystems. *Nat Rev Genet*, 7:510–523, 2006.

[33] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, 74:5088–5090, 1977.

Acknowledgments

First and foremost, I am very grateful to Prof. Dr. Jens Stoye and Prof. Dr. Alfred Pühler for their support, scientific advice, and interest on my work. I would also like to thank the 'Graduiertenkolleg Bioinformatik' of the Faculty of Technology at Bielefeld University for granting me a scholarship that made possible my work on the exciting field of computational biology. For their support and the opportunity to participate in large-scale, international genome research projects, I am grateful to Dr. Alexander Goesmann and Dr. Folker Meyer. Special thanks go to Dr. Robert Edwards and Dr. Forest Rohwer for their inspiration, fruitful discussions, and for giving me the chance to join fascinating metagenomic studies of natural microbial communities. Also parts of this work were conducted during a research visit to the laboratories headed by Robert Edwards and Forest Rohwer, located at San Diego State University, San Diego, CA.

I would like to express my gratitude to all members of the Bioinformatics Resource Facility for the great atmosphere, their support during my work, and their patience, particularly to Naryttza Diaz, Michael Dondrup, Heiko Neuweger, Achim Neumann, Torsten Kasch, Ralf Nolte, and Peter Serocka. For their hospitality and for giving me the opportunity to participate in the effort to 'Annotate a Thousand Genomes' and to visit the Argonne National Laboratory, I feel grateful to Dr. Ross Overbeek, Dr. Veronika Vonstein, and Prof. Dr. Rick Stevens. Furthermore, many thanks go to the International Graduate School in Bioinformatics and Genome Research and its executive director, Dr. Dirk Evers, for their support in this project. Finally, I wish to thank all the people who have helped with fruitful discussions, critical comments, and ideas on the subject, particularly Dr. Alice McHardy, Prof. Dr. Robert Giegerich, Dr. Jörn Kalinowski, Dr. Elisabeth Dinsdale, Juniorprof. Dr. Tim Nattkemper, Dr. Scott Kelley, and Gordon Pusch.

Bielefeld, May 2007

Lutz Krause

ERKLÄRUNG

Ich, Lutz Krause, erkläre hiermit, dass ich die Dissertation selbständig erarbeitet und keine anderen als die in der Dissertation angegebenen Hilfsmittel benutzt habe.

Bielefeld, den 12. Mai 2007

Lutz Krause