GENOME SIGNATURE BASED SEQUENCE COMPARISON FOR TAXONOMIC ASSIGNMENT AND TREE INFERENCE

Author

Kaustubh Raosaheb Patil

Dissertation

for obtaining the degree

of a Doctor of the Natural Sciences (Dr. rer. nat.)

of the Natural-technical Faculties

of the Saarland University

Saarbrücken

2013

SEQUENZVERGLEICH MIT HILFE DER GENOMSIGNATUR FÜR DIE TAXONOMISCHE EINORDNUNG VON SEQUENZEN UND DAS LERNEN TAXONOMISCHER BÄUME

Autor

Kaustubh Raosaheb Patil

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften (Dr. rer. nat.)

der Naturwissenschaftlich-Technischen Fakultäten

der Universität des Saarlandes

Saarbrücken

2013

Tag des Kolloquiums:	29.05.2013
Dekan:	Prof. Dr. Mark Groves
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Hans-Peter Lenhof
Berichterstatter:	Prof. Dr. Alice Carolyn McHardy
	Prof. Dr. Thomas Lengauer, Ph.D.
Beisitzer:	Dr. Nico Pfeifer

EIDESSTATTLICHE VERSICHERUNG

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt ubernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, den 31-05-2013

Kaustubh Raosaheb Patil

ABSTRACT

In this work we consider the use of the genome signature for two important bioinformatics problems; the taxonomic assignment of metagenome sequences and tree inference from whole genomes. We look at those problems from a sequence comparison point of view and propose machine learning based methods as solutions. For the first problem, we propose a novel method based on structural support vector machines that can directly predict paths in a tree implied by evolutionary relationships between taxa. The method is based on an ensemble strategy to predict highly specific assignments for varying length sequences arising from metagenome projects. Through controlled experimental analyses on simulated and real data sets we show the benefits of our method under realistic conditions.

For the task of genome tree inference we propose a metric learning method. Based on the assumption that for different groups of prokaryotes, as defined by their phylogeny, genomic or ecological properties, different oligonucleotide weights can be more informative, our method learns group-specific distance metrics. We show that, indeed, it is possible to learn specific distance metrics that provide improved genome trees for the groups.

In the outlook, we expect that for the addressed problems the work of this thesis will complement and in some cases even outperform alignment-based sequence comparison at a considerably reduced computational cost, allowing it to keep up with advancements in sequencing technologies.

KURZFASSUNG

In dieser Arbeit wird die Verwendung der Genomsignatur für zwei wichtige bioinformatische Probleme untersucht. Diese sind zum einen die taxonomische Einordnung von Sequenzen aus Metagenomexperimenten und zum anderen das Lernen eines taxonomischen Baums aus verschiedenen ganzen Genomen. Diese beiden Probleme werden aus dem Blickwinkel der Sequenzanalyse betrachtet und Verfahren des maschinellen Lernens werden als Lösungsansätze vorgeschlagen. Für die Lösung des ersten Problems schlagen wir eine neue Methode vor, die auf strukturellen Support Vektor Maschinen beruht und direkt Pfade in einem Baum vorhersagen kann, der auf den evolutionären Ähnlichkeiten der Taxa beruht. Die Methode basiert auf einer Ensemble Strategie, um sehr genaue Zuweisungen für Sequenzen verschiedener Länge, die in Metagenomprojekten gemessen wurden, vorherzusagen. Wir zeigen die Vorteile unserer Methode auf simulierten sowie auf experimentellen Daten.

Für das zweite Problem, bei dem ein taxonomischer Baum, basierend auf der genetischen Sequenz gelernt werden soll, schlagen wir eine Methode vor, die eine Metrik lernt. Die Annahme, auf der diese Methode beruht, ist, dass für verschiedene Gruppen von Prokaryoten unterschiedliche Gewichtungen der Oligonukleotidvorkommen notwendig sind, weswegen eine gruppenspezifische Metrik gelernt wird. Die Gruppen können dabei aufgrund ihrer phylogenetischen Beziehungen oder ökologischer sowie genomischer Merkmale bestimmt sein. Wir zeigen in unserer Analyse, dass es hierdurch möglich ist, spezifische Metriken zu lernen, die zu besseren Bäumen für diese Gruppen führen.

Wir erwarten, dass unsere hier vorgestellten Arbeiten für die bearbeiteten Probleme Alignment-basierte Ansätze ergänzen und teilweise sogar überbieten können, wobei unsere Lösungen deutlich weniger Rechenzeit benötigen und damit mit dem rasanten Wachstum im Sequenzierbereich schritthalten können.

ACKNOWLEDGEMENTS

This work would not have been possible without support of a number of people and unfortunately it is not possible to mention all of them.

First of all I would like to thank my supervisor Prof. Dr. Alice Carolyn McHardy for her continuous support, understanding and encouragement. I would also like to thank Prof. Dr. Dr. Thomas Lengauer for his support. The interesting discussions with the members of IRG1 (now AlgBio at HHU) and D3 was always inspirational. Especially, I would like to mention Lars Steinbrück, Sebastian Konietzny, Christina Tusche of IRG1 and Lars Feuerbach, Ingolf Sommer, Jasmina Bogojeska and Nico Pfeifer of D3 and Krzysztof Templin from D2. I am also indebted to Joachim Buech, George Friedrich (MPI) and Klaus Dieter-Baer (HHU) for excellent technical support without which this work would not have been possible.

On a more personal note, I thank my friends for their support, especially all the friends I met in Saarbrücken who made my stay interesting and enjoyable. Last but not least, I am grateful to my family for their continuous support and understanding.

CONTENTS

TABLE INDEX

FIGURE INDEX

1 BACKGROUND

In this chapter we will lay out the background for the work in this thesis and provide necessary notations and definitions. Particularly we will briefly discuss the biological background and motivations. Although most of the work is computational in nature, biological background is provided in order to justify the methods and to motivate the computational work. Note that this is not meant to be an exhaustive account of the related fields. Topics that are not relevant to this work are not discussed.

This work exclusively deals with DNA sequences of prokaryotic origin; therefore, we will start by describing those in sections 1.1 and 1.2. In section 1.3 we will describe metagenomics. Section 1.4 introduces sequence comparison including the genome signature paradigm followed by the challenge of data overload due to advances in sequencing technologies in section 1.5. In section 1.6 we provide overview of machine learning techniques followed by a brief description of the addressed problems.

The mathematical notations used in this thesis follow the following convention; scalar variables will be denoted using small italic letters, vectors will be denoted using small bold non-italic letters and matrices will be denoted using capital bold non-italic letters. Vector and matrix elements will be denoted using non-bold italic letters along with a subscript. The transpose of a vector is denoted using the superscript T.

Several definitions, terms and concepts in this thesis have been taken from other sources as I believe that they cannot be described in a better way. They are indicated with the sign ▷ and the sources are cited. Some of those are modified to match the convention and notation used in this thesis. Some of the frequently used short forms are;

- **Glossary(NCBI) 2002:** Glossary – The NCBI Handbook – NCBI Bookshelf.
- **Metagenomics(NCBI) 2006:** Metagenomics – NCBI Bookshelf.
- **Glossary(Genome):** Genome Glossary – Human Genome Project Information.
- **Glossary(Systematics):** Palaeos – Systematics, Taxonomy, and Phylogeny: Glossary

1.1 DNA AND MOLECULAR EVOLUTION

All known living organisms use genetic material as means to store information and transfer it to next generation underpinning unity of life at a molecular level. Most of the organisms (both unicellular and multicellular) the genetic material used is the deoxyribonucleic acid (DNA) with an exception of quasi-life viruses that use ribonucleic acid (RNA).

▷ **DNA** (Glossary(NCBI) 2002)

> *Deoxyribonucleic acid is the chemical inside the nucleus of a cell that carries the genetic instructions for making living organisms.*

DNA is made of four nucleotide bases (or bases); adenine (A), guanine (G), cytosine (C) and thymine (T). While bases A and G are purines, C and T are pyrimidines. These bases are attached to the backbone structure made out of sugars and phosphate bonds (Levene 1919). The DNA molecule is a double helix structure made of two complementary polymers in which A pairs with T and C pairs with G creating hydrogen bonds resulting in base pairs (bp) (Watson & Crick 1953). This A-T and C-G pairing is called the Watson-Crick base pairing. Thus a DNA molecule can be considered and analyzed using one or more possible structures, including the primary structure which is a base sequence, the secondary structure describing interactions between bases and strands and the tertiary structure describing location of atoms in space. In this work we consider DNA in its primary structure; that is a string made of four nucleotides A, C, G and T. Hereafter all references to a sequence mean a DNA sequence unless otherwise specified.

▷ **DNA sequence** (Glossary(Genome))

The relative order of base pairs, whether in a DNA fragment, gene, chromosome, or an entire genome.

▷ **Gene** (Glossary(Genome))

The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule).

▷ **Genome** (Glossary(Genome))

All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs.

▷ **Chromosome** (Glossary(Genome))

The self-replicating genetic structure of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular, and the entire genome is carried on one chromosome.

▷ **Phenotype** (Glossary(Genome))

The physical characteristics of an organism or the presence of a disease that may or may not be genetic.

DNA is the carrier of genetic information in which genes are information encoding and hereditary units. The central dogma of molecular biology dictates that a gene is transferred into ribonucleic acids (RNA) which is then further translated into proteins that carry out the actual phenotypic functions (Crick et al. 1961). This also implies that the information flows from the DNA to the exterior; consequently the environment affects the DNA only indirectly. This has been disputed and recent understandings in epigenetic modifications and inheritance questions the central dogma (Koonin 2012). We will not discuss this further as the findings in this thesis are not directly affected by whether the central dogma is accepted or refuted.

Before we describe molecular evolution, let us first briefly consider evolution in general. The theory of evolution consists of two mechanisms; descent with modification and natural selection. It was put forward in 1859 by Charles Darwin in his book "On the origin of species by means of natural selection, or the preservation of favored races in the struggle of life" (Darwin 1859). In this book he described how heritable variations combined with natural selection results in survival of the fittest and consequently over a large span of geological time can give rise to the observed biological diversity. As these variations are normally rather small the process of evolution gives rise to a tree-like structure which Darwin depicted in the sole illustration in his book. With the advances in sequencing technologies the evolutionary changes could be studied at a molecular level. Consequently, all known life on the Earth can be represented as a tree depicting evolutionary relationships (Ciccarelli et al. 2006), implying that life originated from a common ancestor. Though the validity of such a tree, especially for prokaryotes, has been questioned (Doolittle 1999, 2000; Bapteste et al. 2009).

▷ **Phylogenetic tree** (Glossary(Systematics))

A branching tree-like, diagrammatic representation of the evolutionary relationships and patterns of branching in the history of the organisms being considered.

▷ **Mutation** (Glossary(Genome))

Any heritable change in DNA sequence.

Each cell contains long structures of DNA called chromosomes which are duplicated during cell division with each cell acquiring its own copy. This process of duplication is not perfect and might cause one of three types of errors; substitution – replacing one type of base by other, deletion – removal of a base and insertion – inserting a new base in the sequence. These errors are called point mutations and lead to novel genotypes. These mutations are either eliminated or become fixed in the genome depending upon whether they are deleterious or advantageous to fitness with respect to natural selection acting upon the phenotypes due to them (Rocha 2008). Alternatively neutral mutations, with no effect on the fitness, can get fixed due to random genetic drift. Furthermore, insertion or deletion of long stretches of DNA can occur by acquiring or removal of transposable elements such as plasmids. Those changes lead to novel genotypes, which can lead to changes in the phenotype as dictated by the central dogma. The phenotypes with an adaptive advantage, for instance efficient utilization of nutrients, reproduce more, in turn increasing the representation of successful genotypes in the population. Alternatively less fit phenotypes reproduce less, thus reducing representation of respective genotypes. Those evolutionary processes can lead to the creation of new species with generations of changes and selection causing the genotypes to be quite different than the one they originated from.

Furthermore, environment can influence genomic features, such as its nucleotide and/or amino acid composition (Foerstner et al. 2005; Willenbrock et al. 2006; Bohlin, Skjerve & Ussery 2009) and physiological structure, either by imposing selective forces or by creating mechanistic mutational biases that in turn can lead to speciation (Orr & Smith 1998; Cohan & Koeppel 2008). Although prokaryotes reproduce asexually, they can recombine within and across lineages. It is generally agreed that the evolution of prokaryotic species is facilitated by

a combination of point mutations and horizontal transfer. Albeit the nucleotide composition pattern is constant within species and varies across species forming the basis of the genome signature paradigm discussed in section 1.4.2.

1.2 PROKARYOTES

The invention of the microscope in the 19th century led to the discovery of the existence of microorganisms. The advent of technologies has rapidly advanced our knowledge about their ubiquitous nature and astonishing phenotypic and molecular diversity. Prokaryotes are single celled ubiquitous microorganisms that lack a cell nucleus. Phylogenetically they make two known domains of life; bacteria and Achaea (Figure 1.1). They show astonishing diversity in habitats and metabolic capabilities making up a large portion of the Earth's biomass. Prokaryotes affect the ecosystem and our own health in many ways. They are a part of the important processes in the ecosystem, such as photosynthesis and nitrogen fixation, cycling of nutrients and production and consumption of organic matter. Prokaryotes, along with other microorganisms like viruses, inhibit various internal and external body parts of higher organisms including human beings and are important for health. Therefore, study of microorganisms is vital not only for understanding of life and ecosystems but also for applied biological sciences such as agriculture and health.

▷ **Genetic marker** (Glossary(Genome))

A gene or other identifiable portion of DNA whose inheritance can be followed.

An easy to ask but difficult to answer question about the prokaryotes is what is the biodiversity of an environment or in other words "how many different species are there in a given environment?" Attempts have been made to, at least partly, answer this question at local and global scales using numerical (Curtis, Sloan & Scannell 2002; Ward 2002) and phylogenetic techniques (Hugenholtz, Goebel & Pace 1998; Hugenholtz 2002). The former has provided an estimate that the entire bacterial diversity of the sea to be about 2 x 10^6 and that of a ton of soil to be 4 x 10^6 different taxa.

At the genome level prokaryotes show high diversity in genome sizes and nucleotide compositional but, remarkably, they all have high coding density with approximately one gene per kilobase (kb), which is not true for eukaryotes (Casjens 1998; Bentley & Parkhill 2004). This high coding density has an important implication on the compositional homogeneity of genomes. Culture independent sequencing (discussed in section 1.3) has greatly contributed towards our understanding of this immense genetic and functional diversity.

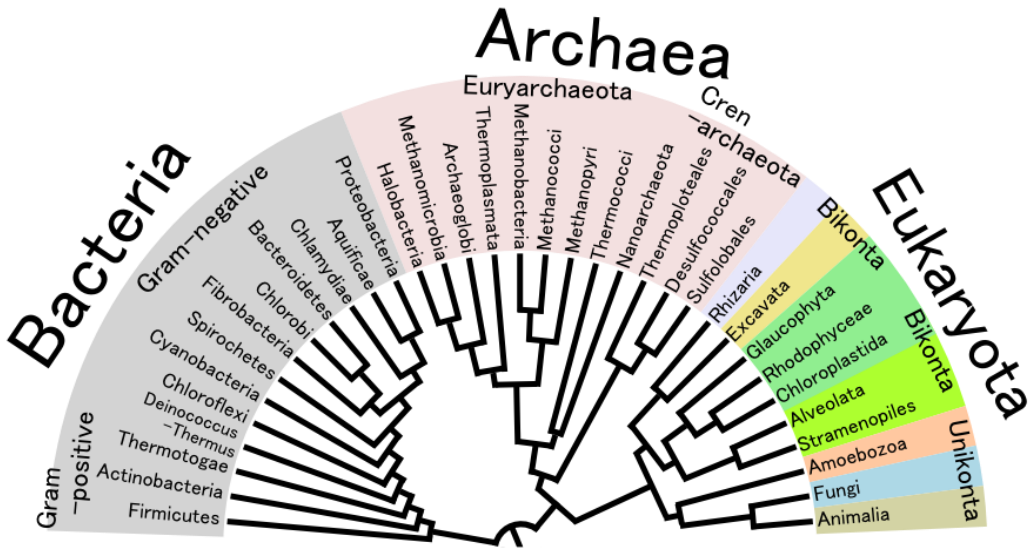


Figure 1.1. Phylogenetic tree showing the diversity of prokaryotes, compared to eukaryotes. From Wikipedia (http://en.wikipedia.org/wiki/File:Phylogenetic_Tree_of_Life.png).

1.3 METAGENOMICS

Genomic studies have advanced our knowledge about molecular basis of life in an unprecedented manner; however, they have some limitations. Sequencing the genome of an organism with traditional methods requires cloning of the entire genome. This is not always possible as majority of the microbes are difficult, if not impossible, to culture in laboratory conditions due to their complex interaction with other species in the community they live in and the environment. Consequently, the uncultivable microbes were and still are largely underrepresented in the molecular databases. This has limited our understanding of the microbial diversity, function and interactions with each other and with the environment. Based on phylogenetic marker gene analyses, these "unknowns" are estimated to represent about 99% of the microbial diversity (Handelsman et al. 1998; Hugenholtz et al. 1998; Hugenholtz 2002; Handelsman 2004).

▷ **Metagenomics** (Metagenomics(NCBI) 2006)

> *Metagenomics is the functional and sequence-based analysis of the collective microbial genomes that are contained in an environmental sample. The word metagenomics describes "the notion of analysis of a collection of similar but not identical items, as in a meta-analysis, which is an analysis of analyses" (Handelsman, Microbiol Mol Biol Rev. 2004).*

Handelsman and colleagues (Handelsman et al. 1998) proposed direct cloning of the collective genomes followed by functional analysis of uncultured soil microbes. By directly extracting DNA from soil and cloning it into readily culturable *Escherichia coli* (*E. coli*), they performed screening for novel chemical products. This opened up a door into the untapped diversity of uncultivable microorganisms. Further progress was made by use of random shotgun sequencing (Tyson et al. 2004; Venter et al. 2004). Numerous metagenomic studies have provided a wealth of information about the structure and function of the communities residing in diverse ecological niches such as the Saragasso sea (Venter et al. 2004), acid mine drainage

(Tyson et al. 2004), sludge processing plant (Garcia Martin et al. 2006) and human and animal gut microbiota (Gill et al. 2006; Turnbaugh et al. 2006; Warnecke et al. 2007; Pope et al. 2010; Turnbaugh et al. 2010; Pope et al. 2012). Such studies not only provide insights into the ecosystems, but also facilitate progress in medicine and biotechnology by identifying genes and enzymes that are drug targets and improve processes such as biomass degradation.

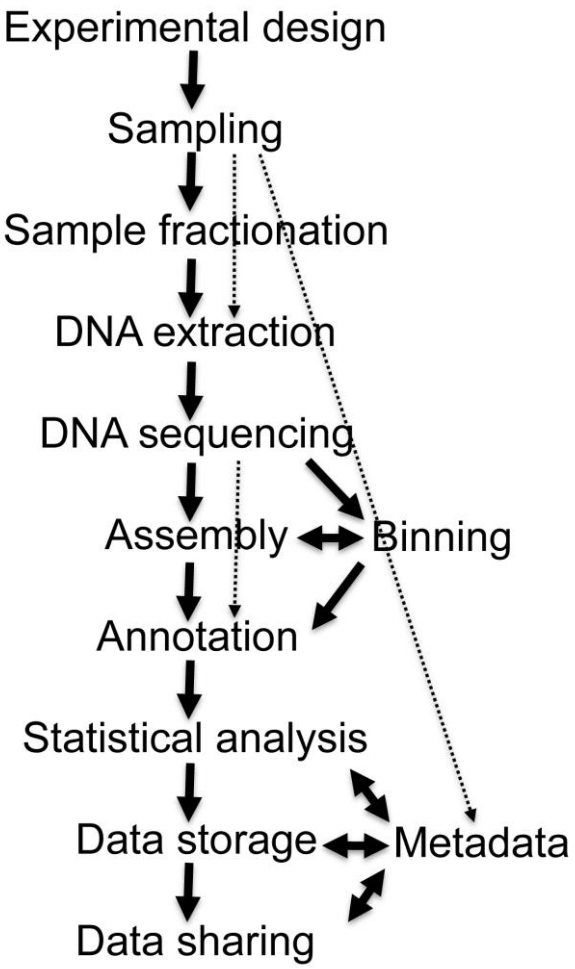


Figure 1.2. Flow diagram of typical metagenome projects. Dashed arrows indicate steps that can be omitted. From (Thomas, Gilbert & Meyer 2012)**.**

The flow-diagram of a typical metagenome project is depicted in Figure 1.2 (Thomas et al. 2012). The output from the DNA sequencing stage are nucleotide sequences (reads) representing the DNA content of the collection of microbes in the sample. Therefore, these studies are often called "community genomics", "environmental genomics" (as sequences for a group of organisms residing in an environment can be obtained) or "metagenomics". These reads can vary in length approximately from 50 bp to 1000 bp depending upon the technology (Table 1.1) and can be subsequently assembled into contigs based on their overlaps. The contigs can be further optionally grouped into scaffolds (or supercontigs) using the paired-end information between the reads in different contigs and the roughly known length between them (Figure 1.3). Note that the scaffolds normally contain unknown sequences of roughly known lengths (gaps) generally indicated by repeating the letter 'N' along the known lengths.

We will refer to all of them simply as sequences. See (Pop, Kosack & Salzberg 2004) for details on the shotgun sequencing and the assembly process.

▷ **Contig** (Metagenomics(NCBI) 2006)

A non-redundant sequence formed by joining, based on sequence overlap, one or more smaller sequences. There should be no gaps.

▷ **Scaffold** (Metagenomics(NCBI) 2006)

A non-redundant sequence formed by joining one or more contig sequences. A sequence overlap is not required to form a scaffold. Typically, a scaffold contains one or more gaps.

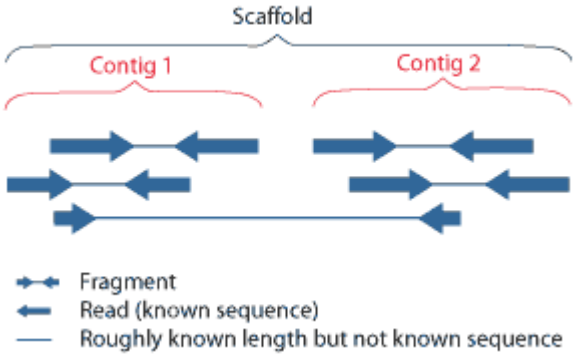


Figure 1.3. The whole-genome shotgun assembly procedure. From JGI Genome Portal (http://genome.jgi.doe.gov/help/scaffolds.html).

As these sequences typically originate from a collection of genomes belonging to the organisms in the community, it is necessary to group the contigs that belong together either at the genome level or at some other higher level, this process is referred to as "binning" (Metagenomics(NRC) 2007; Mavromatis et al. 2007) (Figure 1.2). Binning can be achieved by assigning taxonomic affiliations to the contigs, which is an approach we take in this study. See (Thomas et al. 2012) for a more detailed discussion on metagenomics and involved computational methods.

1.4 SEQUENCE COMPARISON

▷ **Ortholog** (Glossary(NCBI) 2002)

Orthology describes genes in different species that derive from a single ancestral gene in the last common ancestor of the respective species.

▷ **Homologous** (Glossary(NCBI) 2002)

The term refers to similarity attributable to descent from a common ancestor.

Comparison of genomic DNA sequences lies at the center stage in the post-genomic era molecular biology and also of this thesis. Hereafter we will refer to a DNA sequence simply as sequence. The goal of sequence comparison is to identify structural, function or evolutionary similarity between sequences. The basic assumption is that similarity in genomic sequences

reflects higher level similarity. The methods proposed in this work are concerned with the use of sequence comparison in order to quantify evolutionary relatedness between the corresponding organisms. Thus, the sequence similarity reflects evolutionarily relatedness.

Two conceptually different methods are often used to compare genomic sequences; alignment-based and alignment-free methods. Alignment methods, such as the basic local alignment search tool (BLAST) (Altschul et al. 1990), are used to identify orthologs from different taxa based on sequence similarity which subsequently can be analyzed with standard phylogenetic inference methods to infer their evolutionary relationships.

1.4.1 ALIGNMENT-BASED COMPARISON

Two sequences are aligned in order to quantify their identity which in turn can reflect their homology. Two sequences are said to be homologous if they share a common ancestry (Koonin 2005). Such a pair-wise alignment can be either global or local depending upon whether the similarity is considered across the full extent or some regions of the sequences, respectively. Alignments with gaps, representing deletions or insertions, cause the number of possible alignments to grow exponentially with sequence length. Therefore dynamic programming based algorithms were proposed; for example the Needleman and Wunsch algorithm (Needleman & Wunsch 1970) for global alignment and the Smith-Waterman algorithm (Smith & Waterman 1981) for local alignment. Biologically speaking, there is only one true, but unknown alignment between two sequences. In order to find the most plausible alignment the matches, mismatches and gaps in alternative alignments are scored based on scoring matrices and gap penalties and the best alignment is chosen based on the resulting overall scores.

Computational time can be a bottleneck when one wants to compare a query sequence with a database of target sequences. With the growing size of sequence databases an exhaustive search demands a massive amount of time. Basic Local Alignment Search Tool (BLAST) is a heuristic version of the Smith-Waterman algorithm developed for fast database searches. Given a query sequence it scans the database for likely matches before performing alignments consequently reducing search time. Details on BLAST can be found in (Altschul et al. 1990; Pertsemlidis & Fondon 2001).

There are two major shortcomings of alignment-based methods: (i) alignment methods cannot be applied to sequences that are not well conserved across taxa and thus have no orthologs and (ii) they are computationally expensive. Alignment-based similarity is restricted to homologous sequences and cannot be directly applied to sequences with low homology or complete genomes. Furthermore, pair-wise sequence alignment incurs a computational bottleneck because of the $O(N^2)$ asymptotic time and space requirement, where N is the maximum of the lengths of the two sequences being aligned. This makes alignment based algorithms a poor choice for large scale data analyses. Algorithms to compare genomic sequences without alignment were, therefore, proposed, however they tend to be less accurate than alignment-based methods in some settings (Vinga and Almeida 2003; Höhl and Ragan 2007; Reinert et al. 2009). Alignment-free methods utilize the “genome signature”, the evolutionary signal that is contained in the oligonucleotide composition of microbial genomes (Blaisdell 1986; Karlin and Burge 1995).

1.4.2 ALIGNMENT-FREE COMPARISON

In order to address the problems associated with alignment-based comparison, alignment-free methods were proposed. Alignment-free methods primarily rely upon the composition of the sequences in terms of their constituent subsequences. Therefore, knowledge of whole genome or homology is not necessary for alignment-free comparison, as it is not required for the matching subsequences to be contiguous, which is a prerequisite for sequence alignment. Furthermore, the computational complexity of alignment-free comparison is O(N), in contract to the $O(N^2)$ complexity of alignment-based comparison, making it an attractive choice for large scale analyses.

▷ **Oligonucleotide** (Glossary(Genome))

A molecule usually composed of 25 or fewer nucleotides.

Alignment-free comparison stems from the observation that prokaryotic genomes are homogenous in oligonucleotide composition (Rolfe & Meselson 1959; Sueoka 1961a, 1961b; Burge, Campbell & Karlin 1992; Karlin 1994; Karlin & Cardon 1994; Bohlin, Skjerve & Ussery 2008; Blaisdell 1986), meaning that the base composition is invariable for long stretches of sequences within a genome. Due to this characteristic nature the dinucleotide composition is called the genome signature (Campbell, Mrazek & Karlin 1999). Furthermore, alignment-free comparison in general considers a sequence as a continuous unit of information rather than a group of genes (Rocha, Viari & Danchin 1998). Formally a genome signature is expected to exhibit several desirable properties as listed below, in the order of their essentiality.

- **Species-specificity** – a signature should be similar within species and vary across species. This is an essential property for a valid signature.
- **Pervasiveness** – the species-specificity of a signature should pervade the entire genome. This property is essential if the signature is meant to be used for arbitrary segments of a genome.
- **Phylogenetic signal** – distance between signatures should be in accordance with the phylogenetic distance between the corresponding organisms. This property is essential whether evolutionary comparative analyses should be performed.

An excellent review of genome signature along with associated methods and applications can be found in (Vinga & Almeida 2003). Given a sequence several different signatures can be derived; some are discussed in the following sections. In the following, the function fr denotes the frequency of an oligonucleotide assuming that a DNA sequence to calculate the frequency from is given. While the nucleotides are generally denoted using the corresponding capital letter, for example the frequency of cytosine as $\mathrm{fr}(\mathrm{C})$, the oligonucleotides are denoted using place-holders, for example *vxyz* denotes a tetranucleotide.

GC-CONTENT

By analyzing the amounts of nucleotides present in DNA sequences Erwin Chargaff discovered two rules which are known as Chargaff's first and second parity rules, which are essentially

rules of symmetry. Chargaff's first parity rule says that for a double stranded DNA the proportion of A equals that of T and the proportion of C equals that of G (Chargaff 1950). Chargaff's second parity rule extends the first parity rule for sufficiently long (>100 kb) single strands of DNA and is applicable for mononucleotides and oligonucleotides (Rudner, Karkas & Chargaff 1968). While the first rule is a direct consequence of the Watson-Crick base pairing in double-helix structure of DNA (Watson & Crick 1953) the origin and reasons for the second rule are not completely understood (Albrecht-Buehler 2006).

In 1951 Chargaff proposed that the GC-content with respect to total nucleotide counts is species-specific (Eq. 1.1), that is it is constant within a species and varies across species (Chargaff 1951). This has been termed as Chargaff's "GC rule" (Forsdyke & Mortimer 2000).

$$\%\mathrm{GC} = \frac{\mathrm{fr(G)} + \mathrm{fr(C)}}{\mathrm{fr(A)} + \mathrm{fr(C)} + \mathrm{fr(G)} + \mathrm{f(T)}} \qquad \textbf{Eq. 1.1}$$

It has been suggested that this genomic GC-content is related to phylogeny (Sueoka 1961b, 1962; Schildkraut et al. 1962). GC-content, although informative, does not have enough resolution (Sandberg et al. 2003) and is a confounding factor in phylogenetic analyses (Mooers & Holmes 2000; Takahashi, Kryukov & Saitou 2009).

OLIGONUCLEOTIDE SIGNATURE

Similarly to the Chargaff's GC rule, the species-specificity of dinucleotide frequency normalized with the frequency of constituent bases (relative abundances) was established biochemically (Josse, Kaiser & Kornberg 1961; Swartz, Kornberg & Trautner 1962). They observed that the dinucleotide frequencies are non-random, that is their frequency differed from chance expectation, and the relative abundances are different for different species. Those experiments were devised to confirm the Watson-Crick base-pairing and the dinucleotide signature was a side product.

Availability of DNA sequences and advances in information technology allowed computational analyses and further strengthened this idea (Muto & Osawa 1987; Burge et al. 1992). These computational studies established that for long segments of DNA (approximately 50 kb) the dinucleotide relative abundance is species-specific. The dinucleotide relative abundance is defined as the odds-ratio where the numerator is the observed frequency and the denominator represents the expected frequency of the dinucleotide assuming the bases are independently and identically distributed over the sequence, which is a zero-order Markov assumption (Almagor 1983).

$$\rho^*(xy) = \frac{\mathrm{fr}^*(xy)}{\mathrm{fr}^*(x)\,\mathrm{fr}^*(y)} \qquad \textbf{Eq. 1.2}$$

Here $\mathrm{fr}^*(x)$ denotes frequency of an oligonucleotide x on both strands, computed as average frequency of x and its reverse complement. Thus the relative abundance ratio measures the deviation of the observed value from the expected value, causing it to be higher for overrepresented dinucleotides and lower for underrepresented dinucleotides. The species

specificity of this signature was established with the observation that for different species different dinucleotides are over and underrepresented.

Karlin and colleagues also proposed a distance metric to calculate distance between the relative abundances. The corresponding δ^* distance is show below a general form.

$$\delta^*(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{i=1}^{p}\left|\rho^*(x_i) - \rho^*(y_i)\right| \qquad \textbf{Eq. 1.3}$$

Using this distance metric they were able to show that the distances between relative abundances are in accordance with the phylogenetic distance, in other words, the genome signature contains phylogenetic signal. This property has been successfully used by Karlin and colleagues (Karlin & Cardon 1994; Karlin & Burge 1995; Karlin, Mrazek & Campbell 1997; Karlin, Campbell & Mrazek 1998; Campbell et al. 1999) and others (Hao & Qi 2003; Pride et al. 2003; Qi, Wang & Hao 2004b; Sims et al. 2009; Takahashi et al. 2009) to elucidate evolutionary relationship between closely related species based on genome signature. The δ^* distance between whole genome dinucleotide abundances correlates weakly, albeit significantly, with 16S rDNA similarity and strongly with DNA-DNA hybridization values (Coenye & Vandamme 2004).

The signature concept was extended to higher order oligonucleotides, often accompanied by higher order Markov model to calculate the expected frequency, which can show a stronger specificity (Bohlin et al. 2008). In general for a fixed length k the signature over an alphabet Σ is a $|\Sigma|^k$ dimensional vector. In the case of DNA sequences the alphabet is the nucleotides $\Sigma=\{A,C,G,T\}$. The similarity or dissimilarity between sequences is then measured in this oligonucleotide space, allowing use of standard machine learning techniques. Such a representation has been termed “spectrum kernel” and can optionally allow mismatches or gaps (Leslie, Eskin & Noble 2002). The short sub-strings are referred to as oligonucleotides, k-mers, k-tuples or n-grams. We use these interchangeably. Analysis using this representation is also referred to as composition-based analysis or alignment-free analysis as we will refer to it. Similar representation can be derived for protein sequences but it is not discussed as this work focuses upon nucleotide sequences.

Genome signatures have been extensively used to detect laterally transferred DNA (Karlin et al. 1997; Karlin 1998; Pride & Blaser 2002; Dufraigne et al. 2005), inference of evolutionary relationships (Karlin et al. 1997, 1998; Pride et al. 2003; Sims et al. 2009; Xu & Hao 2009) amongst other applications. Hereafter we will refer to oligonucleotide genome signature as genome signature or simply as signature.

ORIGIN AND MAINTENANCE OF GENOME SIGNATURE

The highly variable nucleotide composition of prokaryotes has been observed for a long time (Sueoka 1961b, 1962; Andersson & Sharp 1996), though its origin and maintenance is still not completely understood. In this section some plausible explanations are reviewed.

Two types of evolutionary explanations have been proposed to explain the variation in nucleotide content across prokaryotic species; mutational biases and selective forces. While the former is based on the observation that the GC content in prokaryotes varies from 25% to

75%, suggesting that mutational differences, for example due to differences in DNA replication and repair machinery, play an important role. Furthermore mutational pressure differs in replication strands causing a skew in relative amount of G versus C nucleotide frequencies (McLean, Wolfe & Devine 1998; Lobry & Sueoka 2002). Those observations are consistent with the hypothesis that differences in mutational pressures are responsible for the observed differences in genomic nucleotide content. It was suggested that context dependent mutations, such as CG suppression, can result in some dinucleotides being preferentially generated (Karlin et al. 1997).

Another explanation attributes the observed species specificity of genome signature to selective forces. Many studies have suggested a link between various genomic features and environmental factors such as; exposure to UV is a selective pressure towards high GC content (Singer & Ames 1970), nitrogen fixing aerobes have higher GC content than the ones from the same genus that do not fix nitrogen (McEwan, Gatherer & McEwan 1998), habitat (Rocha & Danchin 2002; Moran, McCutcheon & Nakabachi 2008; Mann & Chen 2010; Botzman & Margalit 2011), optimal growth temperature (Musto et al. 2004; Basak, Mandal & Ghosh 2005; Musto et al. 2005; Kirzhner et al. 2007a; Zeldovich, Berezovsky & Shakhnovich 2007) (see (Hurst & Merchant 2001; Marashi & Ghalanbor 2004; Wang, Susko & Roger 2006) for contrary view), Aerobiosis (Naya et al. 2002; Kirzhner et al. 2007a) and combined effects of phylogenetic and environmental factors (Foerstner et al. 2005; Bohlin et al. 2009; Rudi 2009). Taken together, those findings suggest that genomic nucleotide content contains traces of environmental adaptations, implying the latter being a causative agent.

The pervasiveness of the genome signatures suggests that forces acting on larger stretches of DNA might be involved. That said, to a certain extent signatures may vary within genomes. This intra-genomic variation can be attributed to two different mechanisms. The redundancy of the genetic code allows use of synonymous codons and many organisms show non-random usage. The preferred use of some codons can be due to mutational pressure (Chen et al. 2004) or to adapting the expressional efficiency and accuracy of highly expressed genes (Ikemura 1985; Karlin & Mrazek 2000; Supek et al. 2010; McHardy et al. 2004). Another important source of intra-genomic variation is lateral gene transfer (Koonin, Makarova & Aravind 2001) which causes compositional heterogeneity by introducing foreign DNA. Considering that both sources cause local heterogeneity we ignore them in this work.

GENOME SIGNATURE SETTINGS

At least three parameters need to be set in order to derive genome signatures from sequences and compare them; length of the oligonucleotides, a normalization strategy and a distance metric to compare signatures. All of those choices are vital for the task at hand and are discussed below.

Too short oligonucleotides might not be suitable due to a weaker signal. On the other hand the dimension of the signature vector increases exponentially with the oligonucleotide length resulting in a high dimensional space which can also be problematic, as the distance of a vector to its nearest vector approaches the distance to the farthest vector as the dimension grows (Beyer et al. 1999). Such concentration of distances might render the signatures incomparable. Therefore, a proper choice of oligonucleotide length is necessary for obtaining good results.

Often oligonucleotides between lengths two and ten are chosen in practice, in general, longer oligonucleotides showing stronger species-specificity signal but incur a higher computational cost (Bohlin et al. 2008, 2010). Furthermore, it is known that different oligonucleotide lengths work better for different organisms or groups of organisms (Mrazek 2009). Often oligonucleotides with length between four and six are chosen as they offer a good compromise between signal strength and computational efficiency. Recently a database containing frequencies of oligonucleotides of lengths one to ten has been created (Kryukov et al. 2012). Such databases will eliminate the redundant enumeration of oligonucleotides, further reducing computational requirements.

There are two reasons why one might want to normalize the raw oligonucleotide counts. Firstly, to be able to compare signatures derived from sequences of different lengths. Secondly, to remove biases due to constituent oligonucleotides in order to improve the underlying signal. In the first case, it is sufficient to normalize by the sequence length or total number of oligonucleotides which is a popular choice. In the second case, a count is normalized using the expected count computed using constituent shorter oligonucleotides under a Markov assumption. This can be problematic, particularly for short sequences as the expected count might not be a reliable estimate.

Following the notation used in (Mrazek 2009) we will denote each genomic signature with a pattern *lknm*, where *l* and *n* are place holders for the oligonucleotide length denoted by *k* and the length of oligonucleotides used for normalization denoted by *m*, respectively. As a special case we will use L to denote normalization using the number of nucleotides in a sequence which in turn will be generally represented as $|\mathrm{N}|$ for a nucleotide sequence N. Thus, for example, the tetranucleotide signature normalized using sequence length is denoted as l4nL and normalization using base frequencies is denoted as l4n1. The notation is optionally followed by the alphabet used (e.g. “ry”) if an alphabet other than nucleotide was used.

Each element of a tetranucleotide signature vector normalized using the length for a DNA sequence N is defined as;

$$\rho^{l4nL}_{vxyz|\mathrm{N}} = \frac{\mathrm{fr}(vxyz)}{|\mathrm{N}|} \qquad \textbf{Eq. 1.4}$$

Thus a tetranucleotide signature contains 256 elements (4^4) each corresponding to one tetranucleotide. To take the double stranded nature of the DNA into account, the values of the elements and their corresponding reverse complements ($\mathrm{rev\_comp}$) can be averaged.

$$\rho^{*l4nL}_{vxyz|\mathrm{N}} = \frac{\rho^{l4nL}_{vxyz|\mathrm{N}} + \rho^{l4nL}_{\mathrm{rev\_comp}(vxyz)|\mathrm{N}}}{2} \qquad \textbf{Eq. 1.5}$$

Third choice is the choice of a distance metric to compare genomic signatures. Choices include; the δ* distance due to Karlin and colleagues (see equation Eq. 1.3) (Burge et al. 1992; Karlin et al. 1998), Euclidean distance, cosine distance (Qi, Luo & Hao 2004a), correlation (Pearson or Spearman) based distance (Kirzhner et al. 2002), Mahalanobis distance (Suzuki et al. 2008) and information theoretic distances such as Kullback-Liebler divergence and Jensen-Shanon

divergence (Sims et al. 2009). All of those choices have their own advantages and disadvantages making it difficult to opt for one.

All three choices will be made clear in the respective context.

OTHER SIGNATURES

The signature, or rather the class of signatures, we described above is often referred to as "composition-based" signatures as they represent a sequence as a fixed-length vector derived from oligonucleotide composition. Several other signatures have been proposed and are briefly discussed below.

CODON USAGE

▷ **Codon** (Glossary(NCBI) 2002)

> *Sequence of three nucleotides in DNA or mRNA that specifies a particular amino acid during protein synthesis; also called a triplet. Of the 64 possible codons, 3 are stop codons, which do not specify amino acids.*

The redundancy of the genetic code (many to one association between codons and amino acid) is used in non-random manner by different species (Grantham 1980; Grantham et al. 1980). Grantham's genome hypothesis was based on the observation that genes in a taxonomic group tend to consistently use similar degenerate codons. This qualifies codon usage bias as a genomic signature on the basis of species-specificity and pervasiveness at gene level. However, some consistent inconsistencies are attributed to gene expression level, abundance of corresponding tRNAs and horizontally acquired genes (Ikemura 1985; Sharp & Li 1987). The species-specificity of codon usage was further confirmed by (Wang et al. 2001; Sandberg et al. 2003). Codon usage bias is an attractive choice; however, it requires knowledge of gene boundaries which can be avoided by the use of oligonucleotide based signatures that also pervade non-coding DNA (Campbell et al. 1999).

CHAOS GAME REPRESENTATION (CGR)

Jeffrey (Jeffrey 1990) studied non-randomness of genomic sequences and proposed a visualization technique called CGR which is a 2-dimensional image representation of the sequence. CGR is a generalization of Markov chain processes (Almeida et al. 2001). Deschavanne and colleagues (Deschavanne et al. 1999) drew parallels between CGR and oligonucleotide composition. Later it was realized that for a CGR with resolution is $1/2^k$ and the DNA sequence is much longer than k then the corresponding CGR is completely determined by all the numbers of length k oligonucleotide occurrences (Wang et al. 2005). Therefore, using CGR is, to a large extent, equivalent to using oligonucleotide signatures.

DNA BARCODES

DNA barcodes are short sequences of length 20-25 bp that are present in the genomes of a particular species and are unlikely to be present in the genomes of other species (Stoeckle & Hebert 2008). Barcodes are useful for species identification and classification in an existing taxonomy. However, DNA barcodes are not particularly useful for sequence comparison in

general. Moreover, they do not show genome-wide pervasiveness which is an important requirement for the methods proposed in this work.

OLIGONUCLEOTIDE FREQUENCY DERIVED ERROR GRADIENT (OFDEG)

This signature was proposed by Saeed and Halgamuge (Saeed & Halgamuge 2009) as a single-dimensional genomic signature to extract phylogenetic signals from relatively short DNA sequences. The OFDEG is derived using Euclidean distance (error) between the un-normalized composition vector of a sequence and its sub-sequences of varying lengths. The error decreases with increasing length of the sub-sequences and shows a linear relationship with the sub-sequence length up to certain length. The rate of error reduction within this linear region is the OFDEG value. They showed that OFDEG works as a signature for sequences as short as 200 bp and applied it to the task of taxonomic assignment of metagenome sequences.

1.5 SEQUENCING TECHNOLOGIES AND NEED FOR EFFICIENT METHODS

▷ **Sequencing** (Glossary(Genome))

Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.

▷ **Sequencing technology** (Glossary(Genome))

The instrumentation and procedures used to determine the order of nucleotides in DNA.

The first sequencing technology was developed by Frederick Sanger and colleagues (Sanger & Coulson 1975; Sanger, Nicklen & Coulson 1977) and is known as the “Sanger sequencing” or “chain terminator sequencing”. Post-Sanger sequencing technologies are normally referred to as next generation sequencing (NGS) technologies. NGS technologies produce large amount of sequence data cheaply. Several NGS technologies are commercially available and produce reads of different length, quality and amount. An overview is shown in Table 1.1. Further details on the NGS technologies can be found in reviews (Metzker 2010).

Advancements in genomics, particularly in sequencing technologies, along with the hardware and software aspects of information technologies have fueled rapid development in basic and applied biological sciences. However, the enormous amount of sequence data produced by NGS technologies outperforms the development in computational machinery in terms of processing power and storage (Figure 1.4) and present challenges at various stages of processing and analysis (Kahn 2011).

Sequencing technologies are expected to continue providing improvement in sequence amounts and quality in the future causing data overload. Therefore, one of the biggest challenges is to analyze this large scale data to derive useful information. At the same time it is also important to keep the future developments in mind. Consequently conceptual and methodological development is necessary in order to deliver feasible solutions. The genome signature paradigm (section 1.4.2) provides a sequence comparison framework to devise efficient algorithms that can handle large scale sequence data.

Table 1.1: Throughput and read lengths of different sequencing technologies.

Manufacturer and technology	Length (bp)	Throughput*	Normalized throughput** (Mb/h)	Throughput scale***	Time per run
Solexa/Illumina Sequencing by Synthesis	100-150	300 Gb/8.5 days–600 Gb/11 days	1500-2300	104	8.5 days–11 days
Life Technologies/Applied Biosystems SOLiD	50–75	7 Gb/day–20 Gb/day	300–800	103–104	2 days–7 days
Life Technologies/Ion Torrent	100–200	10 Mb/2 h–1 Gb/2 h	5–500	101–103	2 h
Roche/454 Pyrosequencing	550–1000	450 Mb/10 h–700 Mb/23 h	30–45	102	10 h–23 h
Life Technologies Capillary Sanger sequencing	600–900	690 kb/day–2100 kb/day	0.029–0.088	100	~7 h

*Numbers are based on vendor information: Illumina Inc. (www.illumina.com), Life Technologies (www.lifetechnologies.com), Roche/454 (www.454.com). **Normalized throughput is scaled to a 1-h period and rounded. ***The throughput scale is compared with Life Technologies 3730 Sanger chemistry-based sequencer and shows the ratio of throughput values in terms of order of magnitude. Because lack of information on sequencing statistics or commercial availability, Pacific Biosciences (www.pacificbiosciences.com), Oxford Nanopore Technologies (www.nanoporetech.com) and Helicos Biosciences (www.helicosbio.com) are excluded. From (Dröge & McHardy 2012).

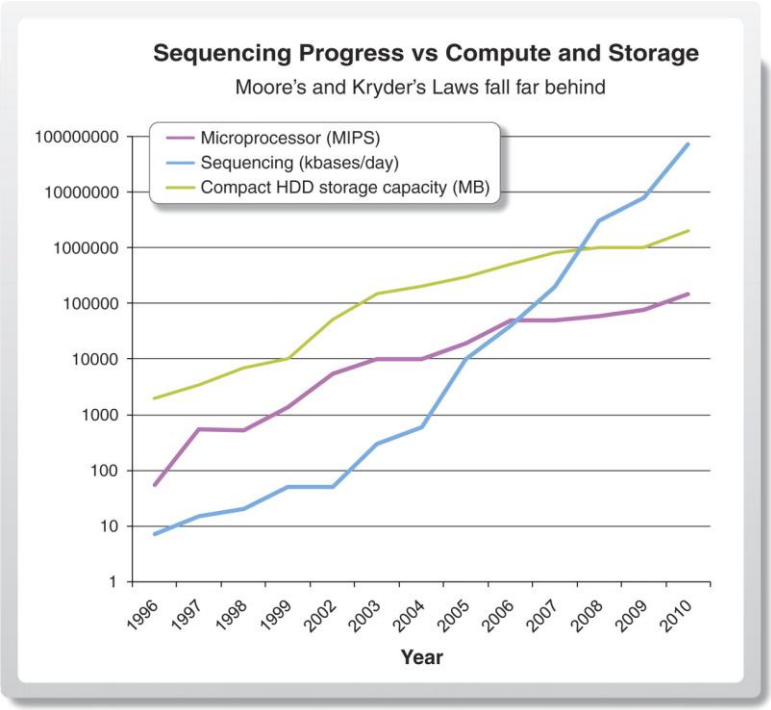


Figure 1.4. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.From (Kahn 2011)**. Reprinted with permission from AAAS.**

1.6 MACHINE LEARNING TECHNIQUES

Having defined the problems addressed in this thesis and described the biological background in the previous sections; this section introduces the machine learning techniques used to solve the corresponding problems.

Machine learning began in the early 1950s and went through many ups and downs as any other scientific disciplines. We will jump straight into defining machine learning. The aim of machine learning is to devise programs, referred to as machines, which learn to perform a task by experience without explicit teaching. A more formal definition was provided by Tom Mitchell.

▷ **Machine learning** (Mitchell 1997)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

The experience needed to learn is normally provided through "training data" which provides the necessary information (independent variables or features or input space) needed to perform the task correctly (dependent variable or output space). The performance measure, as the name says, measures the performance of a learner at the give task. Therefore, these methods learn from empirical data. Depending upon the nature of the training data machine learning methods can be grouped into two categories;

- **Unsupervised (cluster analysis)** – In this case there is no designated output space, often because it is simply not known or is not measured. The training data in this case is said to be unlabeled.
- **Supervised** – In this case the learner has access to the output space. The training data is said to be labeled.

Depending on whether the output space is continuous or discrete a supervised learning problem is said to be either a regression problem or a classification problem, respectively. The nature of the output space further categorizes the classification problems into following three types;

- **Binary** – The output can take one of the two possible values often represented as $\{-1,+1\}$.
- **Multiclass** – The output takes one of the possible m values $\{y_1, y_2, \ldots, y_m\}$.
- **Structured** – This is a generalization of the multiclass problem where the outputs are related to each other in a known structure.

This thesis uses supervised learning methods which are more formally introduced in the following section focusing on the statistical learning theory (Boser, Guyon & Vapnik 1992; Cortes & Vapnik 1995; Vapnik 1995; Hastie, Tibshirani & Friedman 2009).

1.6.1 SUPERVISED LEARNING AND SUPPORT VECTOR MACHINES

The aim of a supervised learning method is to induce a function $f : \mathrm{X} \to \mathrm{Y}$ that maps an input $x \in \mathrm{X}$ to an output $y \in \mathrm{Y}$. Given training data as a finite set of independently and identically distributed (iid) input-output pairs (examples) $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ and a loss

function $\Delta(y, y')$ that quantifies the discrepancy between the correct output y and an output y', the goal of a supervised method is to learn a function such that expected risk over the joint input-output probability distribution $\mathrm{P}(\mathbf{x}, y)$ is minimized;

$$\mathrm{R}(f) = \int_{\mathrm{X,Y}} \Delta(y, f(\mathbf{x}))\mathrm{P}(\mathbf{x}, y)\mathrm{d}\,\mathbf{x}\,\mathrm{d}\,y \qquad \textbf{Eq. 1.6}$$

Due to the unknown probability distribution P the expected risk cannot be computed and has to be induced using a limited training data as the empirical risk;

$$\mathrm{R}_{\mathrm{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} \Delta(y_i, f(\mathbf{x}_i)) \qquad \textbf{Eq. 1.7}$$

▷ **Inducer / induction algorithm** (Kohavi & Provost 1998)

An algorithm that takes as input specific instances and produces a model that generalizes beyond these instances.

The most intuitive loss function for a binary classifier is the 0/1 loss which incurs a penalty of 1 for an incorrect output and no penalty for correct output.

$$\Delta_{0/1}(f(\mathbf{x}), y) = \begin{cases} 0 & \text{if } f(\mathbf{x}) = y \\ 1 & \text{otherwise} \end{cases} \qquad \textbf{Eq. 1.8}$$

This is a step function and hence not differentiable and non-convex. Hence a convex approximation is often used for large margin classifiers, called the hinge loss.

$$\Delta_{+}(f(\mathbf{x}), y) = \max(0, 1 - y' \times f(\mathbf{x})) \qquad \textbf{Eq. 1.9}$$

Here $y' \in \pm 1$ is the true label. We consider function f as a linear hyperplane represented using a vector $\mathbf{w}$ (parameters) with same dimensionality as the input space and an optional bias term b.

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \qquad \textbf{Eq. 1.10}$$

The sign of the function $f(\mathbf{x})$ gives the corresponding predicted output y. The scalar product of two real valued vectors $\mathbf{w}$ and $\mathbf{x}$, $\mathbf{w}^{\mathrm{T}}\mathbf{x}$ is an inner product.

▷ **Inner product** (PlanetMath)

An inner product on a vector space V *over a field* K *(which must be either the field* $\mathbb{R}$ *of real numbers or the field* $\mathbb{C}$ *of complex numbers) is a function* $\langle\cdot,\cdot\rangle$:V×V⟶ K *such that, for all* $a,b \in$ K *and* $\mathbf{x},\mathbf{y},\mathbf{z} \in$ V

1. $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z}\rangle = a\langle \mathbf{x}, \mathbf{z}\rangle + b\langle \mathbf{y}, \mathbf{z}\rangle$
2. $\langle \mathbf{x}, \mathbf{y}\rangle = \overline{\langle \mathbf{y}, \mathbf{x}\rangle}$
3. $\langle \mathbf{x}, \mathbf{x}\rangle \geq 0$, and $\langle \mathbf{x}, \mathbf{x}\rangle = 0$ if and only if $\mathbf{x} = 0$

Inductive algorithms normally estimate the optimal parameters by minimizing risk on the available finite training data, which is the empirical risk. A learned machine (or a fitted model) is then used to predict the output for unseen input data. Therefore, it is important to estimate the predictive capability of a model before employing it. This estimated performance on unseen data is referred to as "generalization performance". Direct minimization of empirical risk can be problematic as it is an ill-posed problem leading to multiple possible solutions and the expected risk might be high even with a low empirical risk (over-fitting). Vapnik proposed finding a hyperplane that is as far away as possible from either of the classes, or in other words a hyperplane with largest margin. Furthermore, a complexity control mechanism is introduced and one often needs to balance two conflicting goals in order to find a generalizable model; empirical risk and the model complexity. This balance forms the basis of the regularization theory and statistical learning theory (Evgeniou et al. 2002). Intuitively the complexity control can be seen as application of Occam's razor where simpler solutions are preferred. Vapnik showed that choosing of a model from a set of models by simultaneously minimizing the empirical risk and maximizing the margin leads to a lower expected risk. Maximizing the margin is equivalent to minimizing the capacity of the machine as defined by the notion of Vapnik-Chervonenkis (VC) dimension providing a probabilistic upper bound on the expected risk. Resulting is the following soft-margin SVM optimization problem for the binary classification task;

Optimization problem SVM_{binary}^{primal}: Given a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{\pm 1\}$

$$\begin{aligned} &\min_{\mathbf{w},b,\xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) \geq 1 - \xi_i \qquad \forall i \end{aligned} \qquad \textbf{Eq. 1.11}$$

Here C is a hyper-parameter that controls the trade-off between the empirical risk and the complexity of the solution. This is known as the "soft margin" SVM as it allows some misclassifications that are penalized using the slack variables ξ. The resulting solution maximizes the margin (distance between the hyper-plane and closest point of each class) around the separating hyper-plane. A dual form of this optimization problem can be derived using Lagrangian multipliers (Vapnik 1995);

Optimization problem SVM_{binary}^{dual}:

$$\begin{aligned} &\max_{0 \leq \boldsymbol{\alpha} \leq C} \sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j \\ &\text{s.t. } \sum_{i=1}^{n}\alpha_i y_i = 0 \qquad \forall i \end{aligned} \qquad \textbf{Eq. 1.12}$$

The solution of this problem results in a set of examples with non-zero weights (α value) which are called support vectors. In the linearly separable case, these are the examples closest to the hyperplane. The primal parameters can be obtained using the following equation;

$$\mathbf{w}^* = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i \qquad \textbf{Eq. 1.13}$$

The prediction function takes the form;

$$f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} \mathbf{w}^* + b = \sum_{i=1}^{n} y_i \left(\alpha_i \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j + b \right) \qquad \textbf{Eq. 1.14}$$

As both the optimization and the prediction functions can be expressed in terms of the inner products between the input examples they can be rewritten using a "kernel function" K over the examples.

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \qquad \textbf{Eq. 1.15}$$

This ability to express both the learning and the inference problems in terms of inner products allows use of any symmetric similarity function that is positive semi-definite satisfying;

$$\mathbf{x}^{\mathrm{T}} \mathbf{M} \mathbf{x} \geq 0 \qquad \mathbf{x} \in \mathbb{R}^n, \ \mathbf{M} \in \mathbb{R}^{n \times n} \qquad \textbf{Eq. 1.16}$$

This assures that the kernel is an inner product between the input examples in some Hilbert space H (feature space) via a mapping $\varphi : \mathrm{X} \rightarrow \mathrm{H}$.

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^{\mathrm{T}} \varphi(\mathbf{x}_j) \qquad \textbf{Eq. 1.17}$$

▷ **Inner product space** (PlanetMath)

An inner product space (or pre-Hilbert space) is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with an inner product $\langle \cdot, \cdot \rangle$.

▷ **Hilbert space** (PlanetMath)

A Hilbert space *is an inner product space which is complete under the* induced *metric.*

Thus similarity between the input examples can be computed in a high dimensional, possibly infinite, feature space without explicit mapping. This "kernel trick", that is a linear solution in the feature space can be non-linear in the input space, is used to solve non-linear classification problems using the linear formulation discussed above. In summary, supervised learning is achieved by identifying a set of parameters such that the expected risk is minimized. A learning method can be generally represented as $L : S, \Theta \rightarrow w$; where Θ is a set of hyper-parameters that are "tuned" for model selection as described below. In the formulation above the hyper-parameter is the regularization constant C. The generalization of the binary SVM to multiclass and structured output will be discussed in section 2.3.1.

1.6.2 Model selection via cross-validation

The choice of hyper-parameters affects the induced model and it is necessary to choose a model with lower expected risk. Cross-validation is a popular technique used for this purpose.

▷ **Model selection** (Hastie et al. 2009)

Estimating the performance of different models in order to choose the best one.

▷ **Cross-validation** (Kohavi & Provost 1998)

A method for estimating the accuracy (or error) of an inducer by dividing the data into k mutually exclusive subsets (the "folds") of approximately equal size. The inducer is trained and tested k times. Each time it is trained on the data set minus a fold and tested on that fold. The accuracy estimate is the average accuracy for the k folds.

The hyper-parameters of a method are varied in order to identify values suitable for data at hand as estimated by the best cross-validation performance. We have used three-fold cross-validation along with a grid search to vary the hyper-parameters in the proposed methods in order to identify optimal models.

1.6.3 METRIC LEARNING

Quantifying similarity or dissimilarity between observations is central to many applications. Often some standard measure is employed for this purpose, for example the Euclidean distance. However, such "off-the-shelf" metric might not be always suitable for the task at hand. Data driven approach can be used to learn a distance metric such that when applied to the target data it produces distances close to the desired distances. We will refer to this problem as metric learning problem.

The Mahalanobis distance metric (Mahalanobis 1936) provides a principled way to represent and learn custom metrics. It is defined as;

$$\mathrm{Mahal}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{y})} \qquad \textbf{Eq. 1.18}$$

Where $\mathbf{x},\mathbf{y} \in \mathbb{R}^p$ are input examples (vectors) and $\mathbf{M} \in \mathbb{R}^{p\times p}$ is a positive semi-definite matrix satisfying Eq. 1.16. Note that the Euclidean distance is a special case of Mahalanobis metric parameterized by an identity matrix. The metric learning problem can then be defined as identification of an adequate matrix $\mathbf{M}$ such that the resulting Mahalanobis distances are close to the desired distances.

As with the supervised classification problem (section 1.6) the goal here is to learn a generalizable metric that can accurately predict the taxonomic distances between new genomes using their genome signatures.

EVOLUTIONARY STRATEGY

As the objective function of the resulting optimization problem is not differentiable, discontinuous and non-convex, gradient based techniques cannot be used. We, therefore, have used evolutionary strategy (ES) (Hansen, Muller & Koumoutsakos 2003) based optimization framework suitable for numerical optimization in this scenario. Evolutionary strategies are based on the concept of natural evolution in the sense that change in a genotype (problem solution) leads to a change in the phenotype (objective function) and

better solutions can be found by recombination and mutation (variation of existing solutions) combined with selection of good solutions produce better solutions over generations (iterations). In a variation of ES the mutation step size for each coordinate of the solution space can be adapted and the correlations amongst them can be accounted for via a covariance matrix. This is called as the Covariance Matrix Adaptation Evolution Strategy (CMA-ES). In other words, the covariance matrix of the distribution is adapted in such a way that the variance is increased in the favorable directions.

1.7 ADDRESSED PROBLEMS

In this thesis we addressed two important bioinformatics problems from the realm of sequence comparison. We rely on the paradigm of genome signature for sequence comparison with the overall aim to achieve good performance with a low computational cost.

1.7.1 TAXONOMIC ASSIGNMENT OF METAGENOME SEQUENCES

▷ **Taxon (plural: taxa)** (Glossary(Systematics))

A group of organisms, considered to be a unit, and which generally has been formally named with a scientific (Latin or Greek) proper name and a rank.

▷ **Rank** (Glossary(Systematics))

The hierarchical level of a supra-specific taxon, according to the Linnaean approach to classification.

▷ **Taxonomy** (Glossary(Systematics))

The field of science convened with discovering, describing, classifying, and naming organisms.

The sequence data generated by metagenomics presents many opportunities to understand the microbial communities and effectively use the knowledge generated. One important and natural question to ask is “who is out there?”. This question can be answered by estimating the taxonomic composition of a metagenome sequence sample. Phylogenetic surveys can answer this question but do not allow us to ask and answer further questions such as “which sequences belong to what taxa?”. This is the taxonomic assignment problem where the goal is to assign taxonomic affiliation to the sequences. Taxonomic assignment allows functional and process-level analysis of the community and possibly genome reconstruction either in whole or in parts. Taxonomic assignments can be obtained by comparing the metagenome sequences with reference sequences with known taxonomic affiliation. In the simplest sense one can assign to a sequence the taxonomic affiliation of its closest match. Indeed, such assignments were performed in the initial metagenome projects (Venter et al. 2004). From the machine learning point of view such a method or its variants are termed a supervised learning techniques (McHardy & Rigoutsos 2007), as they need sequences with known taxonomic affiliation (training data). Unsupervised techniques, on the other hand, are a class of techniques that do not need training data and only use the similarity or dissimilarity between the sequences in a sample to group them. Unsupervised techniques are typically less accurate than supervised techniques when appropriate training data are available.

There are two important challenges in the taxonomic assignment task; a large number of sequences to perform taxonomic assignment on and availability of only partial or no closely related reference data. The oligonucleotide based genome signature paradigm provides a suitable framework that is capable of addressing both challenges. An important thing to clarify is whether the genome signature is applicable to metagenome sequences.

The concept of genome signature was established using analyses of cultivated organisms. As discussed above, two important properties of a genome signature are species-specificity and pervasiveness (genome-wide conservation). Although environmental forces shape nucleotide composition, the genome signature is still prevalent in metagenome sequences (Teeling et al. 2004; Abe et al. 2005) even in extreme environments such as acid mine drainage (Dick et al. 2009). Therefore, genome signature based analyses can be applied to metagenome sequences and several methods have been proposed for taxonomic assignment of metagenome sequences (McHardy et al. 2007; Diaz et al. 2009; Saeed, Tang & Halgamuge 2011) in addition to alignment-based methods (Huson et al. 2007; Krause et al. 2008; Monzoorul Haque et al. 2009; Segata et al. 2012; Sharma et al. 2012). In this thesis we propose a novel method that relies on the genome signature paradigm and uses state-of-the art supervised machine learning methods to achieve good performance with high computational efficiency. Details are provided in chapter 2 and 3.

1.7.2 GENOME TREE INFERENCE

Understanding and inferring evolutionary relationships between organisms is vital. The evolutionary relationships are normally depicted in the form of a phylogenetic tree or a phylogeny. Earlier phylogenies were derived using morphological and physiological characteristics (Orla-Jensen 1909; Stanier & van Niel 1941). Classical examples for bacterial morphological characteristics include cell shape, motility and Gram stain. This clearly poses a problem for microorganisms since it is difficult to identify and characterize morphological features and thus provides a limited resolution for separating different taxa. Later it became clear that morphological and physiological characteristics do not reflect phylogenetic relationships between prokaryotes (Stanier & Van Niel 1962; van Niel 1946). Advent of genomics allowed use of molecular data and thus revolutionized phylogenetic systematics. Molecular sequence based phylogenies are often derived using short molecular sequences such as the ubiquitous small subunit ribosomal RNA genes (16S rRNA and 18S rRNA) that delineated the three domains of life and confirmed the gram stain dichotomy (Woese & Fox 1977; Fox et al. 1980). Although very popular, use of 16S rRNA is not without limitations. As this gene represents only a plausible relationship between organisms and genes are prone to differential evolution rates and horizontal transfer phylogenies inferred using different genes often disagree. To reconstruct the evolutionary history of the organisms many methods were developed that consider several genes (Ciccarelli et al. 2006; Wu & Eisen 2008). These methods rely on multiple sequence alignment of homologous genes.

Following the advent of sequencing technologies a large number of complete genomes became available and it was possible to probe whether evolutionary signals can be found in this rich data source. Traditionally used sequence alignment cannot be directly used for complete genomes as it is only applicable to homologous sequences. Furthermore, genomes

can be non-collinear due to processes of recombination, rearrangement and gene gain and loss. Therefore, several other sources of information are considered, such as gene content, gene order and genomic signature (Delsuc, Brinkmann & Philippe 2005; Snel, Huynen & Dutilh 2005).

▷ **Gene transfer** (Glossary(Genome))

> *Incorporation of new DNA into and organism's cells, usually by a vector such as a modified virus. Used in gene therapy.*

Given evidence of horizontal transfer as a stronger evolutionary mechanism than previously anticipated, the tree-like evolution of prokaryotes is under scrutiny, discussed at length in (Bapteste et al. 2009). We believe that a tree-like representation nonetheless provides practical means to understand the diversity of and relationships between prokaryotes. The usefulness of a tree-like representation is demonstrated by our method for taxonomic assignment of metagenome sequences (see above, Chapter 2). Furthermore, there is a clear distinction between a phylogeny and taxonomy. While phylogeny is meant to describe evolutionary relationships by means of vertical inheritance, taxonomy is a classification system that categorizes organisms into hierarchically organized groups (not necessarily ancestral) along with associated nomenclature conventions (Sneath 1989; Kampfer & Glaeser 2012). In this context the work discussed in Chapter 4 should be viewed as elucidating taxonomic relationships between the organisms which might or might not be evolutionary in nature.

Details for this problem and our solution are provided in chapter 4.

2 PHYLOPYTHIAS FOR TAXONOMIC ASSIGNMENT OF METAGENOME SEQUENCES

Metagenome studies analyze communities of microorganisms from an environment of interest by direct sequencing, thus giving access to uncultivable organisms. A routinely performed step in metagenomic analysis is the taxonomic assignment of the obtained sequences, a procedure known as taxonomic classification. Accurate classification of metagenome samples is a challenging task and depends on the complexity of the microbiome sample, data quality and taxonomic distance to reference genomes. Furthermore, the amount of data produced by next generation sequencing technologies has created a novel challenge namely; there is now a need for fast methods that are scalable for data sets of 500 Mb of sequence in size or more. We present a new taxonomic classification method, PhyloPythiaS, which takes the relationships between taxa into consideration using the structured output prediction paradigm.

2.1 INTRODUCTION

Circumventing the need for isolation and cultivation of individual microbes, metagenomic studies provide insights into the vast and mostly uncultured microbial world. This not only allows the study of microorganisms unreachable by traditional genomics approaches but also facilitates community-level analysis. It has been estimated (Hugenholtz 2002) that 99% of the microbial diversity is uncultured. This produces immense interest in metagenomic studies, with the hope of increasing our knowledge of biodiversity and discovering novel proteins that are of biotechnological or biomedical interest.

Metagenome projects generate a large number of sequencing reads, representing the genetic content of the organismal mixture from the sampled environment. Various computational analyses can be performed on this data; assembly, gene prediction, diversity estimation, and taxonomic assignment some of the common tasks. In the taxonomic assignment problem sequence fragments are assigned to taxonomic units or so-called bins (therefore it is also called as taxonomic binning). The individual bins stand for the species or higher level taxa represented by the populations in the metagenome sample. Taxonomic assignment can be performed either on the reads or on assembled sequence fragments, such as contigs and scaffolds (see sections 1.3 and 1.5).

Three sources have been extensively used to obtain reference taxonomic information for this task; phylogenetic analysis of 16S ribosomal RNA (rRNA) (Woese & Fox 1977), other conserved marker genes (Von Mering et al. 2007; Wu & Eisen 2008) and clade specific marker genes (Segata et al. 2012), similarity searches in sequence databases (Huson et al. 2007; Monzoorul Haque et al. 2009), and sequence similarity in terms of sequence composition, that is using genome signature (McHardy et al. 2007; Diaz et al. 2009; Patil et al. 2011). Marker gene based studies typically assign very few sequences, less than 1% (Hugenholtz 2002). Sequence databases are mainly populated with sequences that are of particular interest, such as biomedical and biotechnological applications (Wu et al. 2009). As sequence similarity searches based on alignment require complete genome sequences, they often fail to identify similar sequences. Sequence composition represents an attractive method for taxonomic assignment

as accurate models can be learned from small amounts of reference sequence (approximately 100 kb), which in some cases can be obtained directly from the sample.

From a methodological perspective, sequence composition-based methods can be categorized as either unsupervised or supervised approaches. A problem which particularly affects unsupervised methods is that the data may be noisy, and influenced by processes unrelated to taxonomic origin such as the environment. For example closely related sequences from different environments; such as farm soil and ocean, differ in their GC content (Foerstner et al. 2005), different lifestyles; such as free living, symbiotic, intercellular and extracellular pathogens, show specific codon usage biases (Willenbrock et al. 2006) and genomic purine composition (A+G) is positively correlated with optimal growth temperature (Zeldovich et al. 2007). Consequently, this may misguide the clustering process towards groupings that might not corroborate with taxonomic origin. Given sufficient amounts of reference sequence, supervised methods can better cope with noisy data by guiding the learning process to focus on the features/examples in a way that confirms with the known class labels.

This chapter presents the design of the PhyloPythiaS method and the associated web server. PhyloPythiaS is a successor to the previously published method PhyloPythia (McHardy et al. 2007) and its name stands for PhyloPythia Structured as it is based on the structured output prediction paradigm. We will describe the associated machine learning techniques in section 2.3 followed by the output and input space and associated choices 2.3.2. In section 2.4 the PhyloPythiaS workflow is presented, in section 2.5 we will present the web server and the chapter ends by showing the advantage of structured output prediction methods in section 2.6.

2.2 EXAMPLES OF DOWNSTREAM ANALYSES

Taxonomic assignment of metagenome sequences facilitates further downstream analyses in turn generating insights into the molecular basis of the biological phenomenon. In order to motivate the work and emphasize the importance of the taxonomic assignment problem, this section provides examples of downstream analyses to gain biological insights in two metagenome projects. Both metagenome samples were analyzed using PhyloPythiaS, in addition to PhyloPythia, achieving high performance as discussed in sections 3.5.4 and 3.5.5. The corresponding samples are described in the section 3.3.2.

Pope and colleagues (Pope et al. 2010) performed compositional and comparative metagenomic analyses of the foregut microbiome of the marsupial; Tammar wallaby (*Macropus eugenii*). The resulting metagenome sequences were taxonomically binned using PhyloPythia. The sequences assigned to a dominant lineage WG-1 from the family Succinivibrionaceae were then used to reconstruct its partial metabolism, devising cultivation-based strategies (Pope et al. 2011). This allowed isolation and characterization of a strain representing the WG-1 lineage, subsequently revealing the microbiological basis for lower methane emissions from macropodids. Taxonomic assignments for this metagenome were also obtained using PhyloPythiaS in order to test the performance of the new method. The results show that both PhyloPythia and PhyloPythiaS performed well and assigned approximately 2.6 Mb to WG-1 with >97% scaffold-contig consistency (described in section

3.2.2). Both methods also showed similarly high performance for other two dominant populations WG-2 and WG-3 (detailed analysis in section 3.5.4).

Another way to use taxonomic assignment is to compare metagenomes in order to identify similarities and differences in their composition, for example taxonomic or genetic, that are potentially associated with a phenotype of interest. Turnbaugh and colleagues (Turnbaugh et al. 2010) performed a study that included comparison of taxonomic bins to identify similarities and differences in deeply sequenced gut microbiomes from monozygotic cotwins. Taxonomic assignment with PhyloPythia identified 25 and 24 genus- and family-level bins were identified in the TS28 and TS29 microbiomes respectively, out of which 22 were common. This taxonomic assignment provided an opportunity for in-depth analysis revealing that *Faecalibacterium* had the highest level of variation, whereas *Methanobrevibacter* had the lowest. The metatransciptome analysis using complementary DNA (cDNA) was then performed respective to the taxonomic assignments, which allowed the authors to calculate the relative expression levels of each bin and gene. This in turn was used to characterize pathways represented by genes with high or low relative expression, which showed that pathways for essential cell processes, e.g. Pyruvate metabolism and Glycolysis, were consistently represented by relatively highly expressed genes. Generic tools have been developed for identification of differentially abundant bins between two or more microbial communities (Huson et al. 2009; Segata et al. 2011). Taxonomic assignment of the metagenome sequence is an essential step prior to such comparative analysis.

2.3 PHYLOPYTHIAS

Building upon PhyloPythia (McHardy et al. 2007) we have developed a new binning method, PhyloPythiaS (Patil et al. 2011; Patil, Roune & McHardy 2012), which uses support vector machine (SVM) based supervised learning method for structured output spaces (Altun, Tsochantaridis & Hofmann 2003; Tsochantaridis et al. 2005; Rousu et al. 2006). Structured output learning exploits a structure which relates different output variables - in this case taxa and their taxonomic relationships as specified by taxonomy - to improve classification performance. Moreover, the structural SVM is based upon the maximum margin principle which gives theoretical generalization guarantees and has also empirically shown good performance. The taxonomic information is obtained from the NCBI taxonomy, which is used to model the evolutionary relationships between taxa or groupings of organisms. Thus, the taxonomic assignment problem becomes a path prediction problem where the output variables (taxa) are organized in a hierarchical structure and the training data consists of oligonucleotide composition of genome fragments of known phylogenetic origin. In the following sections we will first introduce the supervised learning methodology followed by the choice of the input and output spaces.

2.3.1 MACHINE LEARNING TECHNIQUES

STRUCTURED OUTPUT PREDICTION

Structured output prediction is different from binary and multiclass prediction in that the classes are not independent but have some known relationship defined using some structure.

In the present case, this structure is a taxonomy representing the relationships between a set of taxa. In section 1.6.1 we introduced the binary SVM. Two generalizations of the binary classifier have been proposed. The first one extends the classification problem of more than two classes (multiclass) (Crammer & Singer 2001) and the second extends to classification of more than two interdependent classes (structured output) (Altun et al. 2003; Tsochantaridis et al. 2005). Those extensions are introduced next and the link between them is pointed out. For simplicity the discussion is restricted to linear functions and the primal form of the optimization problems, more details can be found in the corresponding references.

MULTICLASS SVM

Many real world problems contain more than two classes and the corresponding classification problem is referred to as multiclass classification. We will denote the output space of m classes as integers $\mathrm{Y} = \{1,2,...,m\}$. In this section we will briefly mention the ideas behind the multiclass SVM. Continuing with the previous notation (section 1.6.1), the learning function, the inputs, the outputs and the parameter vector will be denoted as f, $\mathbf{x}$, y and $\mathbf{w}$, respectively. The bias term b is ignored for simplicity but without loss of generalization.

Two types of methods can be found in literature for the multiclass classification task. The first types of methods decompose the multiclass problem into a set of independent binary classification problems. The most popular strategy is one-versus-all; that is given m classes one first constructs m binary classifiers that separate a particular class from the rest (Crammer & Singer 2001; Rifkin & Klautau 2004). Thus, a different parameter vector is learned for each class $\mathbf{w}_y \in \mathbb{R}^p$. At classification time a new input example is classified by all the classifiers and the class label of the class y yielding the highest positive value is chosen as the output (see section 1.6.1, Eq. 1.10).

$$f(\mathbf{x}) = \underset{y \in \mathrm{Y}}{\operatorname{argmax}}\ \mathbf{w}_y^{\mathrm{T}}\mathbf{x} \qquad \textbf{Eq. 2.1}$$

Other strategies include all-versus-all classification, error correcting codes and defining class structure. Those will not be discussed here and the reader is referred to (Platt, Cristianini & Shawe-Taylor 2000; Pimenta & Gama 2005) and references therein for details.

The second type of techniques can naturally handle multiclass problems, such as nearest neighbor and decision trees (Mitchell 1997; Hastie et al. 2009). Large margin frameworks for construction of a single classifier to handle multiclass problems have been proposed (Vapnik 1998; Weston & Watkins 1999). Crammer and Singer (Crammer & Singer 2001) generalized the notion of margins to multiclass problems and proposed an optimization problem with an efficient algorithms to solve it. Their generalization of the linear binary classifier models the hypotheses space as a matrix $\mathbf{M} \in \mathbb{R}^{m \times p}$ with each row corresponding to a parameter vector for one of the m classes. As the matrix $\mathbf{M}$ can be viewed as stacked parameter vectors (one corresponding to a particular class), , without losing the meaning, we will represent the hypotheses space as a set of m parameter vectors $\mathbf{w}_y \in \mathbb{R}^p$, $\mathrm{y} \in \{1,2,...,m\}$, in order to continue with the notation used in this work. The inference problem is defined similarly to Eq. 2.1.

Further, they proposed a notion of the margin as the difference between the score of the correct row and the maximum of the scores due to one of the other rows (most violating score). The piecewise linear bound on the error for an input vector $\mathbf{x}$ with correct output y is given by;

$$\max_{y' \in Y \setminus y} \mathbf{w}_y^{\mathrm{T}} \mathbf{x} - \mathbf{w}_{y'}^{\mathrm{T}} \mathbf{x} \qquad \textbf{Eq. 2.2}$$

This loss function becomes zero for correct classification and produces a number proportional to the difference between the correct score and the most violating score. The empirical risk in this case is given by;

$$\mathrm{R}_{\mathrm{emp}}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \left[\max_{y' \in Y \setminus y} \mathbf{w}_{y_i}^{\mathrm{T}} \mathbf{x}_i - \mathbf{w}_{y'}^{\mathrm{T}} \mathbf{x}_i \right] \qquad \textbf{Eq. 2.3}$$

Here $\mathbf{w}$ is the concatenation of all the parameter vectors. Defining the norm of the hypothesis space as the norm of the concatenation of all the parameter vectors and using slack variables to allow non-separable data, the optimization problem becomes;

Optimization problem $SVM_{multiclass}^{primal}$**:**

$$\begin{aligned} &\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\ &\text{s.t. } \mathbf{w}_{y_i}^{\mathrm{T}} \mathbf{x}_i - \max_{y' \in Y \setminus y_i} \mathbf{w}_{y'}^{\mathrm{T}} \mathbf{x}_i \geq 1 - \xi_i \quad \forall i \end{aligned} \qquad \textbf{Eq. 2.4}$$

Here C > 0 is a constant controlling trade-off between the empirical risk and the model complexity.

STRUCTURED OUTPUT SVM

The multiclass framework of Crammer and Singer was generalized to incorporate structured output prediction problems (Altun et al. 2003; Tsochantaridis et al. 2004). This framework allows generalization across classes by capturing the common properties of the classes as defined by their interdependencies. This framework can learn over an arbitrary structure among classes, but we will discuss only the special case of hierarchical classification, which is relevant for this work. An important distinction for the structured output paradigm is that an output is a vector instead of a scalar, in the particular case of hierarchical classification each output is a path in the hierarchy with m nodes $\mathbf{y} = [n_1, n_2, .., n_m]$. As before, each input vector is of dimensionality p.

The structured output inference problem is generally defined as;

$$f(\mathbf{x}) = \underset{\mathbf{y} \in Y}{\operatorname{argmax}} \ \mathbf{w}^{\mathrm{T}} \psi(\mathbf{x}, \mathbf{y}) \qquad \textbf{Eq. 2.5}$$

The joint input-output space ψ is defined depending upon the problem at hand, as described below for hierarchical classification. Consider a hierarchy as a set of elements $Z \supseteq Y$ along

with a partial order $\prec$, each path in the hierarchy can be then represented as a vector. Each element of this vector is defined over every element z in Z as following;

$$\lambda_z(\mathbf{y}) = \begin{cases} \beta_{\mathbf{y},z} & \text{if } \mathbf{y} \prec z \text{ or } \mathbf{y} = z \\ 0 & \text{otherwise} \end{cases} \qquad \textbf{Eq. 2.6}$$

Here $\beta_{\mathbf{y},z} \in \mathbb{R}$ defines the similarity between the outputs with respect to the partial order $\prec$. We set $\beta_{\mathbf{y},z}$ to 1, thus obtaining a binary vector for each possible output. In other words, each output (a path in the hierarchy) is represented as a binary vector of size |Z| (number of nodes in the hierarchy) whose each element shows whether a particular node is included in the output or not. Consequently, all the paths containing a node will have a 1 in the corresponding position of the binary representation, indicating the "sharing" between related outputs. Denoting the binary representation for an output $\mathbf{y}$ by $\Lambda(\mathbf{y})$ and an input example $\mathbf{x}$ in the feature space as $\varphi(\mathbf{x})$, the joint input-output space is then defined using the tensor product $\otimes : \mathbb{R}^p \times \mathbb{R}^m \rightarrow \mathbb{R}^{p \times m}$ such that $\mathbf{c} = \mathbf{a} \otimes \mathbf{b}$ $c_{ij} = a_i \times b_j$ as;

$$\psi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) \otimes \Lambda(\mathbf{y}) \qquad \textbf{Eq. 2.7}$$

For example, consider the linear feature map for an input $\mathbf{x}$ defined as $\varphi(\mathbf{x})=\mathbf{x}$ and the binary representation of an output $\mathbf{y}$ as $\Lambda(\mathbf{y})=[1\ 0\ 1\ 1\ 0]$ as a path consisting of three nodes in a hierarchy with five nodes. Then the joint feature space is given by $\psi(\mathbf{x},\mathbf{y})=[\mathbf{x}\ \mathbf{o}\ \mathbf{x}\ \mathbf{x}\ \mathbf{o}]$, where $\mathbf{o}$ is a vector of Zeros of the same length p as the input vector $\mathbf{x}$. Hereafter, we will use the input space feature map function φ as defined above.

This is equivalent to introducing a parameter vector $\mathbf{w}_z \in \mathbb{R}^p$ for every node z in the hierarchy. Thus the complete hypotheses space can be represented as a vector $\mathbf{w} \in \mathbb{R}^{p \times |\mathrm{Z}|}$, which is a concatenation of all $\mathbf{w}_z$ vectors. Note that even though there might not be any input examples directly observed at some paths they use input examples from their children, thus enabling generalization across classes using the compatibility score defined below.

$$\mathrm{F}(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \sum_{z:\mathbf{y} \prec z \text{ or } \mathbf{y}=z} \mathbf{w}_z^{\mathrm{T}} \mathbf{x} \qquad \textbf{Eq. 2.8}$$

Analogous to the Crammer and Singer notion of the margin, a more general functional margin for structured output problems is defined as follows;

$$\gamma(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \psi(\mathbf{x}, \mathbf{y}) - \max_{\mathbf{y}' \in \mathrm{Y} \setminus \mathbf{y}} \mathbf{w}^{\mathrm{T}} \psi(\mathbf{x}, \mathbf{y}') \qquad \textbf{Eq. 2.9}$$

After fixing the minimum functional margin to 1 and penalizing for margin violations the soft-margin optimization problem becomes;

Optimization problem $SVM^{primal}_{structured}$:

$$\min_{\mathbf{w},\,\xi\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } \mathbf{w}^{\mathrm{T}}\,\psi(\mathbf{x}_i,\mathbf{y}_i) - \mathbf{w}^{\mathrm{T}}\,\psi(\mathbf{x}_i,\mathbf{y}'_i) \geq 1 - \frac{\xi_i}{\Delta(\mathbf{y}_i,\mathbf{y}'_i)} \qquad \forall i,\ \forall \mathbf{y}'_i \in \mathrm{Y}\setminus\mathbf{y}_i \qquad \textbf{Eq. 2.10}$$

The constraints require that for each training example the score of the correct output ($\mathbf{y}_i$) must be greater than the score of all incorrect outputs ($\mathbf{y}'_i$) by a margin of 1. There are a large numbers of constraints in this optimization problem $\mathrm{O}(n|\mathrm{Y}|)$which makes it intractable to solve by standard quadratic solvers. As only a small number of these constraints are expected to be active and overlap of information in the joint feature space, an efficient cutting plane algorithm was proposed that guarantees a solution to arbitrary precision by evaluating a polynomial number of constraints. The details of this algorithm are out of the scope of this work and can be found in (Tsochantaridis et al. 2005). The above formulation is the n-slack formulation, since it assigns one slack variable to each training example. A 1-slack formulation of the structural SVM problem was proposed which is computationally more efficient (Joachims, Finley & Yu 2009). The slack-rescaling 1-slack formulation is;

Optimization problem $SVM^{primal,1\text{-}slack}_{structured,slack\text{-}rescaling}$:

$$\min_{\mathbf{w},\,\xi\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi$$
$$\text{s.t.} \frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{y}_i,\mathbf{y}'_i)\left[\mathbf{w}^{\mathrm{T}}\,\psi(\mathbf{x}_i,\mathbf{y}_i) - \mathbf{w}^{\mathrm{T}}\,\psi(\mathbf{x}_i,\mathbf{y}'_i)\right] \geq \frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{y}_i,\mathbf{y}'_i) - \xi, i\in 1..n, \forall \mathbf{y}'_i \in \mathrm{Y} \qquad \textbf{Eq. 2.11}$$

Similarly the margin-rescaling version of the problem is formulated as follows;

Optimization problem $SVM^{primal,1\text{-}slack}_{structured,margin\text{-}rescaling}$:

$$\min_{\mathbf{w},\,\xi\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi$$
$$\text{s.t.} \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{w}^{\mathrm{T}}\,\psi(\mathbf{x}_i,\mathbf{y}_i) - \mathbf{w}^{\mathrm{T}}\,\psi(\mathbf{x}_i,\mathbf{y}'_i)\right] \geq \frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{y}_i,\mathbf{y}'_i) - \xi, i\in 1..n, \forall \mathbf{y}'_i \in \mathrm{Y} \qquad \textbf{Eq. 2.12}$$

The reader is referred to (Joachims et al. 2009) for the duals of those optimization problems. We used duals of the above optimization problems as implemented in the SVMstruct application programming interface available at http://svmlight.joachims.org/ (version 3.10).

2.3.2 Output and input spaces

The output space

Our output space comprises a hierarchical structure representing a set of taxa (nodes) and their taxonomic relationships (edges). Essentially it is a rooted tree. In particular, we use the

taxa and relationships defined by the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/) as the reference. In this structured representation, each possible output corresponds to a valid path in the hierarchy. Each path is encoded as a binary vector of length equal to the number of nodes. In this vector, the elements corresponding to the nodes in the path are set to one and the rest to zero. If an internal node has some training examples assigned to it a miscellaneous terminal child node is added as its child followed by re-assignment of the corresponding training examples to the child node. We used the seven major taxonomic ranks; species, genus, family, order, class, phylum and superkingdom to define the hierarchy.

PATH LOSS

The 0/1 loss used for binary and multiclass problems is not suitable for hierarchical classification as some predictions can be more correct than others. A more suitable loss in this scenario is the path loss. The path loss measures the number of edges on the shortest path between the terminal nodes of two paths (Figure 2.1).

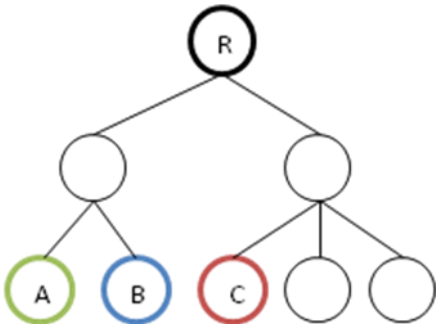


Figure 2.1. The concept of the path loss. For the shown hierarchy, predicting the path from the root node R to the node B is more correct than predicting the path to node C when the correct output is the path to the node A. In this case the path loss for the paths to B and C are 2 and 4, respectively.

As we are dealing with a rooted tree the path loss can be implemented using the depth of the terminal nodes of the corresponding paths and the depth of their lowest common ancestor (LCA);

$$\Delta_{\text{path}}(\mathbf{y},\mathbf{y}') = \text{depth}(\mathbf{y}) + \text{depth}(\mathbf{y}') - 2 \times \text{depth}(\text{LCA}(\mathbf{y},\mathbf{y}')) \qquad \textbf{Eq. 2.13}$$

The path loss can be normalized using the longest path distance in order to restrict maximum loss at one. Other loss functions over a hierarchy can be defined, such as the measure due to (Wu & Palmer 1994); however, we decided to use the path loss for its simplicity and good performance (Cesa-Bianchi, Gentile & Zaniboni 2006; Rousu et al. 2006).

USE OF DYNAMIC PROGRAMMING

Each output in our structured output prediction problem is a path in the taxonomy. Both learning and inference processes depend on the compatibility score (Eq. 2.8), which measures the strength of association between an input-output pair. Let's consider two paths $\mathbf{p}_1$ and $\mathbf{p}_2$ in a hierarchy consisting of nodes $\{n_1, n_2, n_3\}$ and $\{n_1, n_2\}$, respectively. Note that this not the binary representation of the paths, but just an enumeration of constituent nodes. Furthermore

assume the dependency relationship $n_1 \succ n_2 \succ n_3$, saying that n_1 is parent of n_2 and n_2 is parent of n_3. The compatibility score of these paths for a given input vector **x** are given by (see Eq. 2.8);

$$\mathrm{F}(\mathbf{x}, \mathbf{p}_1; \mathbf{w}) = \mathbf{w}_{n_1}^{\mathrm{T}}\mathbf{x} + \mathbf{w}_{n_2}^{\mathrm{T}}\mathbf{x} + \mathbf{w}_{n_3}^{\mathrm{T}}\mathbf{x}$$
$$\mathrm{F}(\mathbf{x}, \mathbf{p}_2; \mathbf{w}) = \mathbf{w}_{n_1}^{\mathrm{T}}\mathbf{x} + \mathbf{w}_{n_2}^{\mathrm{T}}\mathbf{x} \qquad \textbf{Eq. 2.14}$$

Thus calculation of the compatibility score of the whole path needs the compatibility score of its constituent nodes which in turn needs the scores of its ancestors. As a concrete example, the compatibility score for the path $\mathbf{p}_1$ can be rewritten as follows;

$$\mathrm{F}(\mathbf{x}, \mathbf{p}_1; \mathbf{w}) = \mathrm{F}(\mathbf{x}, \mathbf{p}_2; \mathbf{w}) + \mathbf{w}_{n_3}^{\mathrm{T}}\mathbf{x} \qquad \textbf{Eq. 2.15}$$

Therefore a hierarchy can be traversed either in depth-first preorder or in breadth-first level-order to calculate compatibility scores of all the paths. This dynamic programming results in high computational efficiency.

INPUT SPACE SETTINGS

As the input space we use a genome signature defined over a set of oligonucleotides. There are at-least two parameters that have to be defined to get a signature space; the lengths of the oligonucleotides and a normalization strategy of the oligonucleotide frequencies (see section 1.4.2). We first fixed the range of oligonucleotide lengths from four to six, as this choice has been used by various previous works (Teeling et al. 2004; Abe et al. 2005; McHardy et al. 2007). We performed 3-fold cross-validation experiments to identify suitable parameter settings. The cross-validation experiments were performed by varying the regularization constant C (see Eq. 2.11, Eq. 2.12) in the set {0.1, 1, 1000, 10000}. Those experiments were performed on 1332 complete prokaryotic genomes downloaded from NCBI. All the taxa from the superkingdom to the species rank were modeled if at-least three genomes could be assigned to it, which resulted in 401 taxa in total (2 superkingdoms, 21 phyla, 34 classes, 69 orders, 105 families, 105 genera and 65 species).

First we performed cross-validation experiments to identify which normalization to use. Two normalization strategies were tested; sequence length and constituent mononucleotides (zero-order Markov assumption) (see section 1.4.2). Tetranucleotide signatures were calculated using Eq. 2.16 and Eq. 2.17, sequence length and mononucleotide normalization, respectively. Pentanucleotide and hexanucleotide signatures were calculated similarly. Note that the latter incurs a higher computational cost. The oligonucleotide counts are generally not reliable for short fragments and thus we expected the mononucleotide normalization to perform worse on shorter fragments. The mononucleotide normalization performed worse for 1000 bp fragments and comparatively better for 5000 and 50000 bp fragments (Figure 2.2), as expected. It can be observed that with a proper choice of the C parameter the sequence length normalization is able to deliver same cross-validation performance as the mononucleotide normalization. Therefore, we chose sequence length normalization.

$$\rho^{*l4nL}_{abcd|N} = \frac{\mathrm{fr}^*(abcd)}{|N|} \qquad \textbf{Eq. 2.16}$$

$$\rho^{*l4n1}_{abcd|N} = \frac{\mathrm{fr}^*(abcd)}{\mathrm{fr}^*(a)\,\mathrm{fr}^*(b)\,\mathrm{fr}^*(c)\,\mathrm{fr}^*(d)} \qquad \textbf{Eq. 2.17}$$

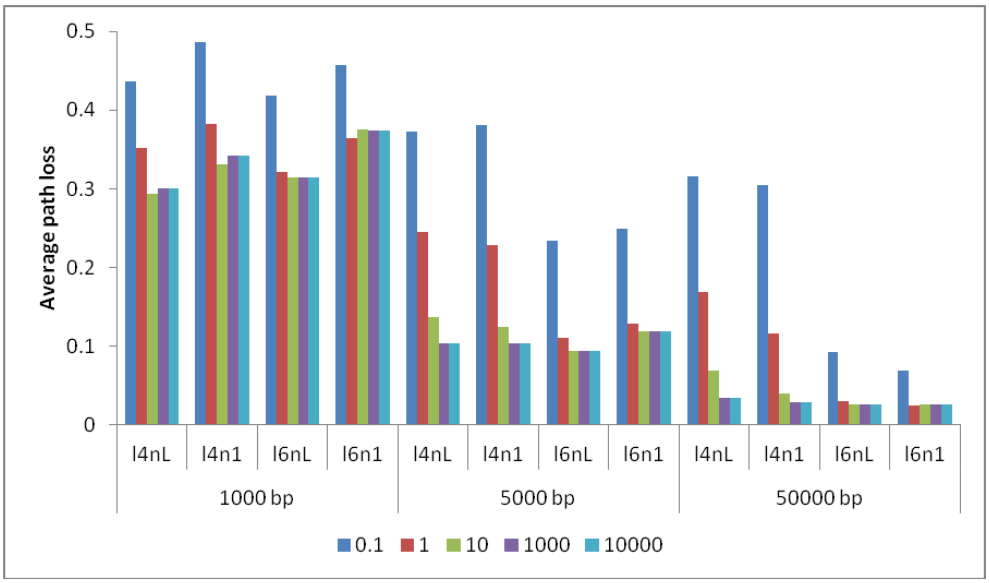


Figure 2.2. Cross-validation experiments to select a normalization strategy. Each bar shows the accuracy for oligonucleotides of selected length and a C value.

The next choice to make is the oligonucleotide length. We investigated the following choices of oligonucleotide lengths (dimensionality in brackets); 4 (256), 5 (1024), 6 (4096) and a concatenation of 4, 5 and 6 (5376). As before; the C parameter was searched in the set {0.1, 1, 1000, 10000}. It was observed that for all three fragment lengths performance improved with increasing dimensionality of the input space (Figure 2.3). Note the trend that cross-validation performance improves for longer fragments, confirming that longer sequence fragments encode a stronger signal. Based on these experiments we chose the input space to be a concatenation of oligonucleotides of lengths 4, 5 and 6 normalized with sequence length.

REGULARIZATION PARAMETER SETTING

The structural SVM problem has a hyper-parameter, the regularization constant C. The choice of this parameter affects the trade-off between the empirical risk and the complexity of the solution. The cross-validation experiments suggest that a C value of 1000 works well for all fragment lengths (Figure 2.3). Therefore, this value was used from then on.

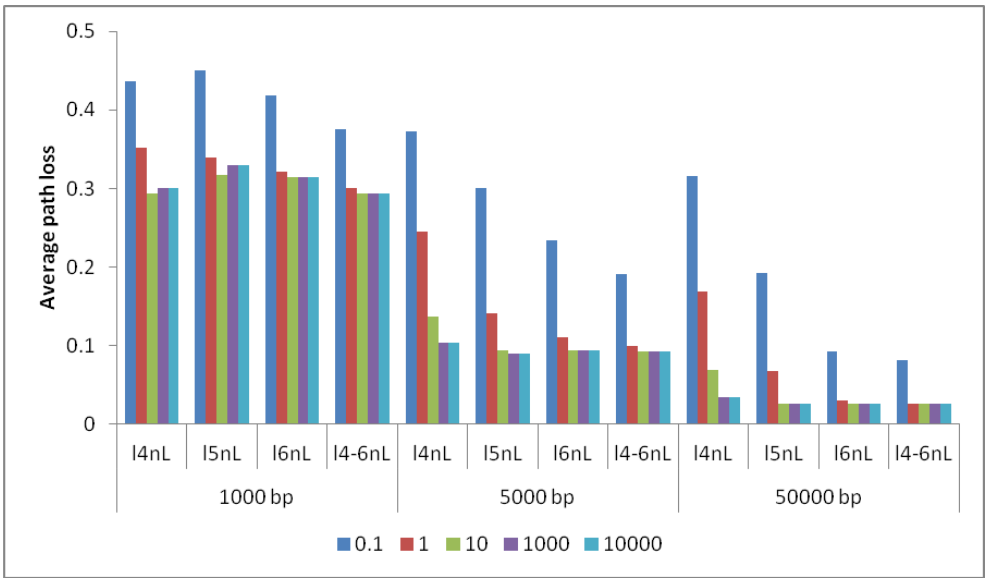


Figure 2.3. Cross-validation experiments to select oligonucleotide lengths. Each bar shows the accuracy for oligonucleotides of selected length and a C value.

2.3.3 Ensemble of classifiers

The prediction problem that we intend to address using structured output SVMs involves potentially differing distributions of training and test data due to the varying lengths of sequences produced in a metagenome project (section 1.3). Following (McHardy et al. 2007) we build six models using six fragment lengths; 1000, 3000, 5000, 10000, 15000 and 50000 bp, to be able to classify sequences of varying lengths typical in metagenome studies. The six structural SVM models induced using genome signatures from each of those fragments comprise a PhyloPythiaS model. Each test example (sequence) is classified with at most three classifiers close to its sequence length or longer. The resulting predictions are then combined using an ensemble strategy. Due to the hierarchical structure of the classes the majority vote ensemble normally used with multi-class techniques is not applicable here. Considering that we would like the predictions to be as specific as possible, that is close to the leaf nodes we devised an ensemble strategy “majority vote lowest node”. In this strategy, first a vote is assigned to each node equal to the number of classifiers predicting a path containing that node. Then, for an ensemble of three classifiers, the nodes with a vote greater than one are traversed in breadth-first order until the corresponding classifiers agree on the predicted path, finally assigning that path as the output (Figure 2.4). In other words, the path on which the majority of the classifiers agree upon is the output of this ensemble strategy.

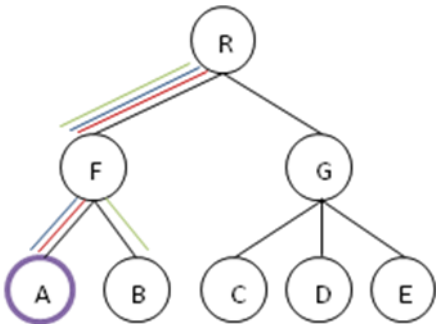


Figure 2.4. The majority vote lowest node ensemble strategy. A colored line adjacent to an edge represents inclusion in the prediction by a classifier, where each color represents a classifier. In this hypothetical case, two out of three classifiers make consistent prediction till node A, thus assigning the path from the root to node A as the output of the ensemble.

2.3.4 GENERIC AND SAMPLE-SPECIFIC MODES

PhyloPythiaS has two different modes of operation – generic and sample-specific.

A generic model is learned from public sequence data from the NCBI along with the corresponding taxonomy. First the taxa are identified for which at least three genomes are available. The reference taxonomy is then completed using the higher level parents of the selected taxa from any of the seven major taxonomic ranks. The generic mode of PhyloPythiaS uses a generic model and is suitable for the analysis of a metagenome sample, if no further information on the sample's taxonomic composition or relevant reference data is available.

Lack of appropriate reference data can cause taxonomic assignments to be either of low resolution (i.e. assignments to high ranking taxa) or inaccurate. There are two reasons why the appropriate reference data might be lacking. Firstly, the vast majority of microbial diversity has not been cultured and sequenced (Hugenholtz 2002), and therefore metagenome samples often represent novel species for which no sequences of closely related organisms are available in public databases. Secondly, although the genomic signature is informative for species and higher-level taxonomic clades (Burge et al. 1992; McHardy & Rigoutsos 2007), it is also known that sequence characteristics are dependent upon environmental factors (Foerstner et al. 2005; Willenbrock et al. 2006). In this case, the genomic signature of the organisms in the metagenome sample can deviate from the genomic signature of the evolutionarily close organisms available in public databases. A sample-specific model (i.e. a model that includes training data from the metagenome sample itself in addition to public data) is better suited in such scenarios. By including sample-specific sequences and taxonomy in the training of SSVM, the dataset shift problem can be reduced (Adams 2010).

Therefore, assignment accuracy can be improved by creation and use of a sample-specific model, which includes clades for the abundant sample population that are inferred from the appropriate reference sequences. A sample-specific model is inferred from the public sequence data combined with sequences with known taxonomic affiliation identified from the metagenome sample (sample-specific sequences), together with a sample-specific taxonomy. The sample-specific sequences along with any other available information, such as the ecology of the sample, can be used to identify the taxa that should be modeled, which along with their

parent taxa makes the sample-specific taxonomy. If a good match between the sample-specific sequences and taxonomy and the taxonomic composition of the metagenome sample is achieved, sample-specific models normally exhibit higher predictive accuracy (discussed in section 3.5), and have improved resolution to low-ranking clades and higher coverage in terms of assigned sequences, compared to a generic model. Normally, accurate assignments can be obtained based on ~100 kb of reference sequence for a modeled sample population. This is possible due to the pervasiveness of the genome signature. Suitable sample-specific training sequences can be obtained from the metagenome sample itself, for example based on sequence homology of the sample sequences to 16S rRNA or other phylogenetic marker genes, or by targeted sequencing of metagenomic fosmid library with such phylogenetic marker genes (Warnecke et al. 2007; Pope et al. 2010).

2.4 THE PHYLOPYTHIAS WORKFLOW

Having described the components of PhyloPythiaS, in this section we will describe the detailed workflow for building a PhyloPythiaS model and making taxonomic assignments using it.

The prerequisites for building a model include DNA sequences, either complete genomes or parts of it, with known taxonomic affiliation and a database of taxonomic relationships. Note that the sequences can come from public databases such as the NCBI GenBank or can be obtained from the metagenome sample itself as sample-specific sequences. Plasmid sequences are omitted if such information is available. The model building starts by cleaning the sequences of undefined characters so that they have minimum effect on the sequence length which is used for normalizing the oligonucleotide counts. For this, contiguous non-ATGC characters longer than the selected oligonucleotide length (k) are substituted by k 'N' characters. This also makes sure that invalid oligonucleotides are not counted. Then the nodes to model are identified based on the taxonomic affiliation of the sequences. While for a generic model the nodes at species or higher level major taxonomic ranks where at least three sequences can be mapped are modeled (this number can be set by the user), for a sample-specific model the nodes to be modeled are defined by the user depending upon the sample composition. This information is then converted into the Newick tree format (nested parentheses format) (Figure 2.5) and retained for later use. All the sequences are then mapped to the lowest possible node in this tree. The sequences are then fragmented into non-overlapping fragments of desired length and an equal number of fragments are selected for each node such that the total number of fragments equals the desired number of training examples set by the user (default value 10,000). Only the nodes where the sequences were mapped are counted. Note that as sample-specific sequences are normally short (~100 kb), they are fragmented into overlapping fragments such that the required number of fragments are generated. If there are more fragments than required then the required number of fragments are randomly sampled stratified with respect to the original sequences, ensuring that every sequence makes equal contribution wherever possible. The genome signature of each of the fragments is then computed as per user defined oligonucleotide length (default 4-6). This set of genome signatures can be represented as a matrix where each row is one signature. We used the sparse matrix representation supported by SVMstruct to store this matrix. This matrix along with the Newick tree is then used to train a SSVM model. Before learning starts every column of the matrix is standardized to have zero mean and standard

deviation of one. The means and standard deviations of each column are retained for later use. To avoid incomplete paths, the tree is modified by adding miscellaneous leaf nodes to the internal nodes if they have sequences assigned to them. The regularization parameter C is by default set to 1000 or it can be also obtained via cross-validation. By repeating this procedure for each of the fragment lengths (default 1, 3, 5, 10, 15 and 50 kb) different SSVM models are obtained which together make a PhyloPythiaS model. Supplementary Figure 1 diagrammatically shows the training process.

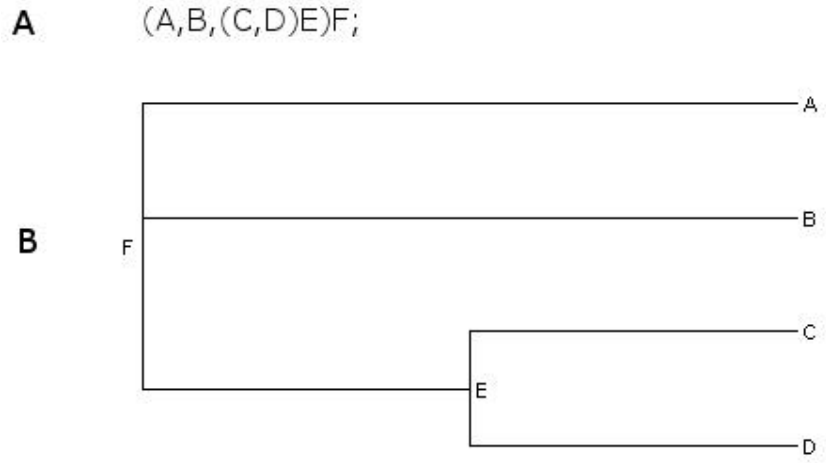


Figure 2.5. A Newick tree example in the nested parentheses format (A) and the corresponding dendrogram visualized using Dendroscope (Huson & Scornavacca 2012) **(B).**

At the prediction time the test sequences are converted into genome signatures using the same settings as used for model building. Each of those genome signatures are then classified with at most three SSVM models built with the fragment lengths closest to that of the test sequence length. The genome signature is standardized using the mean and standard deviation of the corresponding model before running the inference (section 2.3.1, Eq. 2.5). Each SSVM outputs a path in the model taxonomy. If two or more SSVM models were used then the resulting predictions are combined using the majority vote lowest node ensemble strategy (section 2.3.3). This process is repeated for each of the test sequences and all the outputs are written in a file.

2.5 THE PHYLOPYTHIAS WEB SERVER

The PhyloPythiaS software is freely available for non-commercial users and can be installed on a Linux-based machine. For researchers with limited computational resources or who are not familiar with command line usage under Unix/Linux, web servers provide computational resources and a graphical user interface for convenient use. Furthermore, they allow a visual presentation of results for a quick overview and exploration of data sets. Therefore, we implemented a web server that provides the PhyloPythiaS functionality. Several web servers for taxonomic assignment are available, such as the MG-RAST (Meyer et al. 2008), WebCARMA (Gerlach et al. 2009) and the naïve Bayes Classification (NBC) (Rosen, Reichenberger & Rosenfeld 2010) web servers. Our server is unique in that it provides the ability to construct and use sample-specific models, besides enabling assignment with generic models.

As previously described, the web server can be used in two different modes – generic or sample-specific. The generic mode accepts sequences as a multi-FASTA file of up to 100 Mb in size and performs taxonomic assignments using a generic model. The generic model is constructed from prokaryotic genome sequences available at NCBI and models sufficiently covered clades from domain to species level (see Introduction). The sample-specific mode allows the user to specify the clades for a model and upload representative sequences for construction of a user-defined model. In this mode, the user has to provide three files: (1) a tree file: a plain text file with NCBI identifiers for the clades to be modeled or a rooted Newick tree with non-negative integer node names; (2) a sample-specific FASTA file: a multi-FASTA file with sample-specific sequences, where each sequence header must contain a valid node identifier X as “label:X”; and (3) a prediction FASTA file: a multi-FASTA file with the sequences for which taxonomic assignments are to be made. The sample-specific data provided by the user is pooled with the reference data used for generic model to build a model with default parameters as described in previous sections. This model is then used for taxonomic assignment of the test sequences provided in the prediction FASTA file.

The generic and sample-specific models produce output in the same format. The output page shows an assignments table with a maximum of 100 entries, as well as a pie chart and the model taxonomy. The pie chart shows the abundance of the taxa and can be interactively changed to visualize different taxonomic ranks and to display either the number of sequences or number of bases. The taxonomy shows the modeled tree along with the assignment information for each node. The taxonomy can be interactively changed to display either the taxonomic identifiers or the NCBI scientific names.

Such interactivity allows the user to easily visualize the distribution of the assignments over the taxonomy. Every node in the tree contains additional information, such as the number of sequences/bases assigned to the node or its sub-tree. Additionally, a link is provided to obtain the sequences assigned to each node. The assignments can be downloaded, possibly with additional data, or received via email. If the server was invoked in the sample-specific mode then additional assignments on separate data can be obtained using the same model.

Metagenome samples can be larger than the upload limitations of the web server. For this reason, the ability to visualize and download combined assignments from multiple submissions for classification with the same model is provided. One uploads a large sample in the form of multiple non-overlapping FASTA files, each as a different process, and retains the corresponding process identifiers. Once all the processes are finished, the process identifiers can then be provided to the ‘multiplex-sample’ utility, which combines the predictions from all processes and generates visualizations and download files.

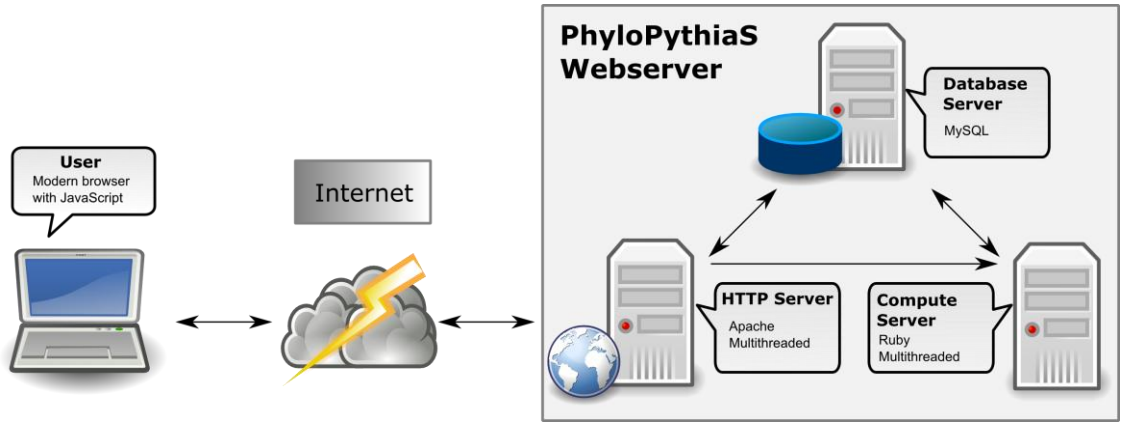


Figure 2.6. Schematic representation of the PhyloPythiaS web server implementation. Arrows represent the direction of communication.

The web server consists of multiple components (Figure 2.6). The web interface is implemented in PHP and JavaScript, and runs on an Apache server. The visualization and help routines are implemented in JavaScript using the Dojo toolkit (http://dojotoolkit.org/). The computational routines for the backend are written in the Ruby scripting language (http://www.ruby-lang.org/) embedded inside an XMLRPC server. These routines pre-process every job to create the necessary files and then invoke binaries compiled from C code (for oligonucleotide feature generation and SSVM). A relational database based on MySQL is used to store the uploaded data, results and configuration. The jobs are processed in the same order they enter the database. The jobs and any associated data are deleted 30 days after their finishing time. The user does not need to register for using the web server, and job identification and result retrieval is done using a unique identifier assigned to every job at the submission time. By default, one processor each is reserved for the generic and the sample specific mode. This can be changed by the administrators in case of large number of pending jobs and depending upon availability of resources.

2.6 COMPARISON WITH FLAT TECHNIQUES

Machine learning based prediction techniques that consider classes independently are known as flat methods. Flat methods are normally faster to learn than the methods that take structure between classes into account (structured methods for short). Furthermore, it has been shown that structured methods for hierarchical classification of documents into genre can perform poorly for imbalanced hierarchies (Wu, Markert & Sharoff 2010). Therefore, it is important to assess whether structured methods provide improvement over flat methods for the taxonomic assignment task. We empirically compared two variants of structural SVM (slack rescaling and margin rescaling) with four flat methods (SVMmulticlass, libSVM, kNN and naïve Bayes) using 3-fold cross validation experiments. SVMmulticlass and libSVM are multiclass extensions of SVMs. While SVMmulticlass is an implementation of the Crammer and Singer multiclass SVM as a special case of structural SVM, libSVM uses one-against-one strategy (Hsu & Lin 2002). The kNN is one of the simplest and oldest techniques with native support for multiclass classification (Cover & Hart 1967). Finally, naïve Bayes is a probabilistic classifier that assumes independence between features (Mitchell 1997). Note that all these techniques have been used in the context of taxonomic assignment (see Introduction). The SVMmulticlass

(version 2.20) was obtained from the author's website http://svmlight.joachims.org/svm_multiclass.html, libSVM and naïve Bayes classifiers were from the "e1071" package (version 1.5-24) and kNN from the "class" package (both in R version 2.11.1).

Two types of validation experiments were performed; class-stratified experiment where examples for each class were randomly split into three folds and leave-classes-out experiment where examples for all the classes were randomly divided into three folds. While the class-stratified experiments assess generalization performance when input examples from all the classes are available, the leave-classes-out experiments assess generalization performance when either no or only closely related input examples are available. The cross-validation experiments were repeated 10 times with different random seeds while maintaining same folds for all the methods. Thus, each technique was tested on 30 folds in total.

The experiments were performed on a dataset with 166 classes with a total of 402 nodes in the hierarchy. The classes belonged to the hierarchy at different taxonomic ranks; species (65 classes), genus (60), family (33), order (6) and phylum (2), indicating the imbalanced nature of the hierarchy. The input space used was the l4nL genome signature. As our aim here was to quantify differences between methods with and without hierarchy other signatures were not tested. Five different C parameters were used for SVMs {0.1, 1, 10, 1000, 10000} and the number of nearest neighbors for kNN {1, 2, 3, 4, 5}. For each of the 10 cross-validation experiments the performance on the folds using the hyperparameter with the best cross-validation performance was used. For all the methods, except naïve Bayes, the data was standardized to zero mean and unit variance. We measured accuracy and path loss for both flat and structured methods.

The paired Wilcoxon signed-rank test was used to compare the performance of any two methods on the 30 folds. On the class-stratified cross-validation experiments both SVMmulticlass and libSVM performed significantly better than all other methods, including structured, in terms of accuracy. However, while libSVM performed better than the structured methods in terms of path loss, SVMmulticlass performed significantly worse that the structured methods. Both kNN and naïve Bayes performed significantly worse than all the other methods with kNN performing relatively better. Also note that SVMmulticlass performs better than structured methods ($P<0.05$, Wilcoxon test) on the accuracy but worse ($P<1e-4$, Wilcoxon test) on path loss measure, suggesting that direct minimization of 0/1 error does not necessarily improve path loss performance (Figure 2.7, Supplementary Figure 2).

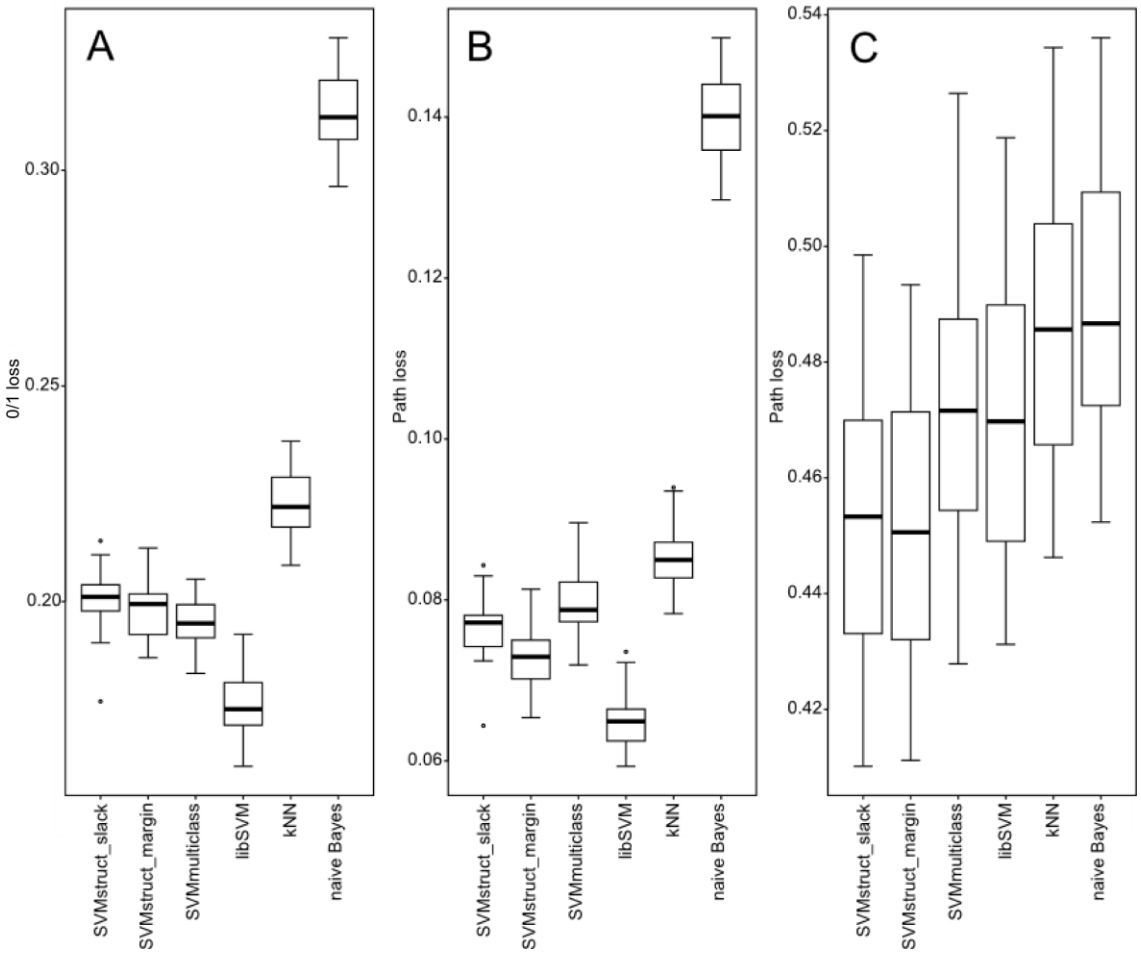


Figure 2.7. Performance of the six machine learning techniques in two cross-validation scenarios. The 0/1 loss and the path loss in class-stratified cross-validation, (A) and (B) respectively. Path loss in leave-classes-out cross-validation (C).

In the case of the leave-classes-out cross-validation, where the complete classes were left out, the structured methods outperformed all flat methods on the path loss measure. Note that in this case the accuracy of all the methods is zero. The margin rescaling formulation performed significantly better than the slack rescaling formulation (P=2.83e-4, Wilcoxon test). Both multiclass SVM methods performed similarly, still outperforming kNN and naïve Bayes (Figure 2.7, Supplementary Figure 2). Note that the average loss for a worst classifier that assigns a label with maximum loss and a random classifier is 0.92 (standard deviation 0.019) and 0.68 (standard deviation 0.0049), respectively, implying that all the techniques tested indeed learn and generalize.

Taken together, those results suggest that structured methods are beneficial when data from the same class is not available for training. We, therefore, expect that as data becomes scarce, structured methods become more beneficial because they can use information from closely related classes (as defined by the hierarchy) analogous to multitask learning scenario (Evgeniou & Pontil 2004).

3 PHYLOPYTHIAS EVALUATION AND APPLICATION

As most of the microorganism diversity is still unknown it is very unlikely that complete genome sequences of the dominant populations are available as a reference for the taxonomic assignment of a metagenome sample. Moreover, there might be varying degree of evolutionary relatedness between the available reference data and the dominant populations. In some cases it is possible to obtain limited amounts of sequence data for the dominant populations which we call sample-specific data. Thus, it is crucial to assess the performance of taxonomic classification methods when limited amounts or no reference data from closely related organisms are available. We, therefore, performed controlled experiments on simulated and real data sets mimicking realistic scenarios. We show that PhyloPythiaS performs well on both simulated and real data and offers a significant improvement in execution time.

3.1 INTRODUCTION

The assignment performance on a metagenome sample depends on a combination of various factors that are either intrinsic or extrinsic to the sample. While intrinsic factors include organismal complexity, data quality and lengths of the sequences, extrinsic factors include the assignment method and availability of closely related reference sequences. In particular, the assignment of short fragments of less than 1000 base pairs is a difficult task (McHardy & Rigoutsos 2007). Although various methods have been developed for this purpose (Krause et al. 2008; Brady & Salzberg 2009; Parks, MacDonald & Beiko 2011), the accuracy remains less than what is achievable for longer fragments. From the extrinsic factors availability of closely related genomes is an important issue and a taxonomic method should be able to cope with the availability of partial genomes or lack of thereof.

Apart from those intrinsic and extrinsic challenges associated with metagenome samples, the large volumes of sequence data generated with next generation sequencing technologies represents a major challenge. In the future, the amount of data generated in metagenome studies will continue to grow, as sequencing comes with further reductions in costs and simultaneous increases in speed (see section 1.5). Therefore, taxonomic assignment methods should be able to cope with this sheer amount of data, while delivering good performance at the same time. However, currently, many binning methods cannot process large data sets in reasonable time.

The design of the PhyloPythiaS method and the evaluation setup was devised while considering the challenges described above. We will first describe the performance measures used in section 3.2, followed by the data sets and taxonomic classification methods in sections 3.3 and 3.4, respectively. The evaluation results are presented in sections 3.5 and 3.6. The chapter is concluded in section 3.7. Note that these analyses were performed at different times and therefore use different reference data, such that the more recent data is a superset of the older data.

3.2 PERFORMANCE MEASURES

3.2.1 SIMULATED DATA SETS

As the correct taxonomic assignment for test fragments is known, evaluation of simulated datasets can be performed using well established performance measures. Here we compute the sensitivity and specificity of assignments, averaged over all the taxa at a fixed taxonomic rank (Baldi & Brunak 2001). The measures are computed for each taxon separately by considering combination of all the other taxa as a different class as shown in the Table 3.1.

Table 3.1. Confusion matrix.

		Predicted class	
		Taxon$_i$	**Taxon$_{-i}$**
Correct class	**Taxon$_i$**	True Positive (*tp*)	False Negative (*fn*)
	Taxon$_{-i}$	False Positive (*fp*)	True Negative (*tn*)

Thus, the average sensitivity, or macro-accuracy, and specificity are defined as follows (Baldi and Brunak 2001; McHardy et al. 2007);

$$\text{specificity} = \frac{1}{n}\sum_{i=1}^{n}\frac{tp_i}{tp_i + fp_i} \qquad \textbf{Eq. 3.1}$$

$$\text{sensitivity} = \frac{1}{n+1}\sum_{i=1}^{n}\frac{tp_i}{tp_i + fn_i} + \frac{tp_{-1}}{tp_{-1} + fn_{-1}} \qquad \textbf{Eq. 3.2}$$

The index -1 denotes items that do not belong to any of the modeled taxa for a given rank. Furthermore, we compute the classification accuracy, which corresponds to the overall number of correctly classified items at a given taxonomic rank. Note that while the macro-accuracy measures the classification accuracy averaged over all classes represented in a test data set, the accuracy measures classification performance for a given data set in a way that every input item contributes equally. This distinction becomes important if the taxa are represented in uneven amounts in a given data set, such as is often the case for metagenomic data, in which case, the overall classification accuracy becomes a more relevant performance measure than the macro-accuracy of assignments.

$$\text{accuracy} = \frac{tp}{tp + fn} \qquad \textbf{Eq. 3.3}$$

Ideally, a method should score well in terms of all measures.

We also have used the average non-normalized path loss (Eq. 2.13) in order to measure the taxonomic distance between the correct and the predicted taxa.

3.2.2 REAL DATA SETS

As for real metagenome samples the correct taxonomic assignment of the fragments is not known, measuring the binning performance on real metagenome samples is a non-trivial task and traditional measures like accuracy, sensitivity and specificity cannot be calculated. We use here an intuitive and informative measure, called “scaffold-contig consistency”, for assessing the binning performance of a method (McHardy et al. 2007). We extended this measure to incorporate contig lengths, as described below.

Consider a metagenome data set for which the reads are assembled into contigs and that a set of contigs are known to jointly originate from a particular genome, based on the mate pair information. This is denoted by their grouping into a scaffold (see Figure 1.3). A taxonomic assignment method is then used to infer the taxonomic assignment of the contigs. The scaffold-contig consistency measures the consistency of the taxonomic assignments for a scaffold in terms of its constituent contig assignments. For this purpose, first the “true” taxonomic assignment for each scaffold is obtained as follows; a scaffold is first labeled with the assignment of one of its constituent contigs with the lowest taxonomic rank. In case there are multiple lowest rank assignments, then the assignment with the longest collective contig length is used. The consistency of scaffold assignments is then measured with respect to this taxonomic label. For each contig of a scaffold, the taxonomic assignment is considered to be consistent if it is either the same or a more general taxonomic assignment with respect to the true taxonomic origin of the scaffold; otherwise it is considered an inconsistent assignment. The percentage of consistently assigned contig base-pairs is the scaffold-contig consistency. The scaffold-contig consistency is then averaged over all the scaffolds with the same assignment, to measure the assignment consistency of a clade. Furthermore, we also calculate the average taxonomic distance of contig assignments in terms of the path distance to the scaffold label as a more fine grained consistency measure. High scaffold-contig consistency is a desirable property for a binning method. For a given data set we use the same reference taxonomy for all the methods for calculating scaffold-contig consistency. Note that the scaffold-contig consistency measure can be tricked by a method that is consistently making wrong predictions, as it can achieve a high performance with this measure. For example, consider a method that assigns same label to all contigs. Such a method will achieve perfect scaffold-contig consistency scores. Nevertheless it is still an informative measure for “honest” methods. Several other performance measures were used for the individual real data sets and will be explained in the respective context.

3.3 DATA SETS

3.3.1 SIMULATED DATA SETS

It is not straightforward enough to incorporate all the complexities present in real metagenome sequence samples such as organismal diversity and novelty in simulated data sets. However, when aware of these limitations, simulated data represent a good starting point for a thorough evaluation. Note that by simulated data we mean sequence fragments that were selected from genomes with known taxonomic affiliation and not simulated sequences or hierarchies. Recently, three simulated datasets of varying complexity were

constructed from fragments of sequenced genomes and used to benchmark the performance of various computational methods, including taxonomic binning techniques (Mavromatis et al. 2007). We use the medium complexity dataset (simMC) from this benchmark collection for evaluation, as well as newly constructed simulated data sets of short fragments.

ACID MINE DRAINAGE DATA SET (SIMMC)

We analyzed the simulated acid mine drainage data set (simMC) (Mavromatis et al. 2007) to evaluate the performance of the different binning methods. We used the data set of contigs assembled with the Arachne assembler, which consist of 7307 contigs of which ~99% come from six strains of three species (two strains each); *Rhodopseudomonas palustris*, *Bradyrhizobium sp. BTAi1* and *Xylella fastidiosa*. The average contig length is 2332 bp. We used the NCBI complete genomes for the training. Controlled sets of genomes were excluded as described in the results section.

SHORT FRAGMENTS DATA SET (SIMSF)

Next generation sequencing technologies yield short reads (~30-1000bp depending upon technology) and produce large amounts of data. It is, therefore, interesting to see, whether it is possible to characterize such short fragments directly without assembly. We simulated short fragments data sets to answer this question. The benchmark data sets were constructed with two constraints: First, the fragments to be characterized should not belong to any of the organisms represented among the reference sequences, as metagenome sample populations are rarely among the available sequenced isolate genomes. Secondly, they should be chosen such that the closest reference genomes are found at different taxonomic ranks, to model different degrees of evolutionary relatedness of metagenome sample populations to available reference sequences. To simulate this set-up, sequences from the NCBI genomes database were used as reference data for model construction. One hundred isolate sequences from the NCBI whole genome shotgun database with no mapping to any of the genera of the reference data were used for testing. Of the latter, 48 belong to a family, 39 to an order and 13 to a class of the reference taxonomy (data not shown). Thus, the test genomes were 'unknown' for PhyloPythiaS; that is not seen during training. Approximately 10,000 non-overlapping fragments of 100, 300, 500, 800 and 1000 bp in length were randomly sampled from the test sequences to create the test sets of varying lengths.

3.3.2 REAL DATA SETS

ACID MINE DRAINAGE METAGENOME SAMPLE (AMD)

The AMD is a well-studied metagenome sample of an acidophilic biofilm community, sequenced with Sanger sequencing technology (Tringe et al. 2005). The AMD community comprises five abundant species: Ferroplasma Types I and II, a Thermoplasmatales species (all Euryarchaeota), and Leptospirillum sp. Group I and II of the phylum Nitrospirae. The test scaffolds for the AMD metagenome were downloaded from the IMG/M portal (http://img.jgi.doe.gov/, taxon object ID 2001200000). These data comprise 1183 scaffolds and ~10.83 Mb of DNA sequence. Draft genome assemblies, comprising 908 scaffolds overall, were

created using sequencing coverage and nucleotide composition for the five populations of the AMD sample; the genome assemblies were then deposited at NCBI (accession numbers CH003520–CH004435). We mapped the AMD scaffolds to these reference assemblies with BLASTN (Altschul et al. 1990) and used the best match in terms of the lowest E-value for each scaffold of the AMD data set as an estimate of its "correct taxonomic affiliation".

TAMMAR WALLABY FOREGUT METAGENOME SAMPLE (TW)

Microbial communities from the gut of the Australian Tammar wallaby (*Macropus eugenii*) were sequenced by Sanger sequencing (Pope et al. 2010) (GenBank accession number ADGC00000000). This sample consists of approximately 13.572 Mb of assembled DNA sequence, with contig lengths varying in length from 438 bp to 27,865 bp (average length 2,276.38 bp). 16S rRNA analysis determined that organisms from the phyla Firmicutes and Bacteroidetes and the gamma-subdivision of Proteobacteria are abundant. This sample contains at least three abundant microbial populations, namely Wallaby gut 1 (WG-1 – a population of an uncultured Succinivibrionaceae bacterium), WG-2 (of a novel deep branching lineage within the Lachnospiraceae) and WG-3 (a novel bacterium of the Erysipelotrichaceae).

HUMAN GUT METAGENOME SAMPLES (HG-TS28 AND HG-TS29)

Two metagenome sequence samples from the gut of two human monozygotic, female twins were obtained by Roche/454 deep sequencing of the total fecal community DNA with 454 Titanium single- and paired-end protocols (Turnbaugh et al. 2010) (referred to as TS28 and TS29). We analyzed approximately 113 Mb and 72 Mb of assembled contig sequences for TS28 and TS29, respectively. Sample-specific training data was obtained with BLASTN homology searches versus a reference database of 118 sequenced gut genomes. Training data was identified based on the following criteria; e-value<10^{-5}, bitscore>50, percent identity>90, percent sequence aligned>90, and total contig length>2 kb. Furthermore, all significant matches were required to originate from the same reference genome.

COW RUMEN METAGENOME SAMPLE (CR)

We furthermore performed taxonomic assignments for 26,042 metagenomic scaffolds (568 Mbp) of a microbial community adherent to switchgrass incubated in a bovine rumen (Hess et al. 2011) with a twofold objective: First, to demonstrate usage of the PhyloPythiaS web server on a large dataset and, second, to verify usability of the method for sequences generated by Illumina sequencing technology. The data was downloaded from the DOE Joint Genome Institute website (ftp://ftp.jgi-psf.org/pub/rnd2/Cow_Rumen/). The majority of the scaffolds were found to have no similarity to sequenced genomes in the original study, suggesting uncharacterized microbes as their origin. Fifteen near-complete 'genome bins' of abundant populations from four orders were identified in the original study from the cow rumen sample, based on analysis of tetranucleotide frequency and assembly information (Hess et al. 2011). We used these genome bins, comprising 466 scaffolds overall, as the correct taxonomic affiliation for comparison with the taxonomic assignments of PhyloPythiaS. The partial genome bins published in the original article are not guaranteed to be entirely correct, nevertheless

they provide a qualitative reference point, as they were generated based on multiple sources of information and verified by human in-depth inspection.

3.3.3 PHYLOPYTHIAS SETTINGS

We used the genome sequences from the NCBI complete genomes repository as reference data for model construction. The output hierarchy was restricted to the taxa to with at least three genomes could be assigned or as defined by the sample-specific data. We built six structural SVM models using different fragment lengths; 1, 3, 5, 10, 15 and 50 kb. For each of these models approximately 10,000 input examples were used equally distributed over all the taxa being modeled. The C value was fixed to 1000 (section 2.3).

3.4 METHODS USED FOR COMPARISON

The following sub-sections give a brief account of taxonomic classification methods used for comparison. All of those methods are based upon supervised machine learning techniques.

3.4.1 PHYLOPYTHIA

PhyloPythia uses patterns of oligonucleotides along with ensemble of hierarchical classifiers combining multi-class SVMs the radial basis function kernel for taxonomic assignment of variable length metagenome sequences (McHardy et al. 2007). PhyloPythia builds a multiclass SVM for each of the domain to genus taxonomic ranks and combines them using a bottom-up approach for hierarchical classification.

3.4.2 PHYMM AND PHYMMBL

Phymm uses interpolated Markov models (IMMs) using sequence composition features for taxonomic classification of metagenome sequences. It was specially designed to classify reads as short as 100 bp. PhymmBL is a hybrid classifier which combines Phymm with BLAST to improve the assignment accuracy (Brady & Salzberg 2009).

The PhymmBL package was obtained from the website http://www.cbcb.umd.edu/software/phymm/. This software by default downloads the NCBI RefSeq and taxonomy data and builds IMMs on the corresponding sequences. The first version of PhymmBL (available when the corresponding analyses were performed) did not allow training on arbitrary sequences, unless some specific conditions on the fasta headers and folder names are met. We, therefore, changed the perl scripts to allow use of arbitrary training data, so that NCBI draft assemblies and sample-specific data could be used.

3.4.3 METAGENOME ANALYZER (MEGAN)

MEGAN requires comparison of the metagenome sequences against databases of known sequences using BLAST or another comparison tool. A lowest common ancestor (LCA) algorithm is then used to assign reads to taxa such that the taxonomical level of the assigned taxon reflects the level of conservation of the sequence. MEGAN offers various parameters for adjustment of the LCA algorithm (Huson et al. 2007).

MEGAN was obtained from the website http://www-ab.informatik.uni-tuebingen.de/software/megan. MEGAN can detect standard NCBI names in the BLAST output, so including sample-specific data was straight forward. We created various BLAST databases; NCBI complete genomes, NCBI draft assemblies and sample-specific data (when available) using the "formatdb" program (available with blast at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/). For sample-specific data, care was taken to include the organism names in the fasta headers before formatting them as a BLAST database, so that MEGAN could detect their taxonomic position. Default MEGAN parameters for LCA were used. Database searches were performed using blastn to appropriate databases using blast alias files. The complexity filter was turned off with option –F "m D" when performing blast searches.

3.4.4 Best BLASTN-hit

This is one of the simplest approaches used for alignment-based taxonomic classification but generally not well suited when closely related genomes are not available. The idea here is to obtain similarities between a test sequence and the reference sequences using BLASTN and then assign the taxonomic affiliation of the reference sequence that yields the highest similarity with the test sequence. The similarity is usually measured using the e-value (Altschul et al. 1990).

We created BLAST databases for appropriate sequences using the "formatdb" command. The metagenome sequences were queried, using "blastn", against this database with default parameters. The resulting blast report was parsed using BioRuby (Goto et al. 2010). Each query sequence is labeled with the taxonomic identifier of the genome with the best hit (lowest e-value). Hits with e-value less than 0.1 were discarded as being insignificant.

3.4.5 Naïve Bayesian classifier (NBC)

The first naïve Bayesian classifier in this context was proposed in (Sandberg et al. 2001). Later many other implementations with some modifications were proposed (Rosen et al. 2010; Parks et al. 2011). In this work we used the web server implementation as described in (Rosen et al. 2010).

We downloaded the assignments provided by the NBC webserver (http://nbc.ece.drexel.edu/ with default N-mer length of 15 and Bacteria/Archaea genomes (accessed in April 2011) and used the "summarized_results.txt" file to extract the sequence headers and species level assignments (columns 1 and 4). These assignments were used for subsequent analysis, for example generating pie charts and predictive performance calculations.

3.5 Results

3.5.1 Acid mine drainage simulated data set

We analyzed the simulated acid mine drainage data set to evaluate the performance of the different binning methods. For this task, complete genomes from NCBI were used as reference data for model training with exception of those genomes used to create the simulated data

set. This corresponds to the unknown genome test setting, in which no training data of the respective populations within a metagenome sample is used. For testing we used the contigs assembled with the Arachne assembler (Mavromatis et al. 2007). The performance of different methods is summarized in Figure 3.1. As can be seen, all methods perform well, overall, on this data set. PhyloPythiaS show very high specificity at all taxonomic ranks, while PhymmBL exhibits highest sensitivity but comparatively lower specificity at lower taxonomic ranks. MEGAN shows average specificity and sensitivity. Overall, PhyloPythiaS is conservative in assignments, tending more towards under-binning than the other methods, which results in a lower overall number of assignments to genus- and family-level clades on this data set.

In order to simulate the effect of varying degree of evolutionary relatedness between the training and test data, we evaluated performance of the different taxonomic classification methods by retaining 100 kb randomly selected contiguous fragments from the three dominant strains each as reference data and removing all genomes of the (1) same genus, (2) same order and (3) same class for the dominant strains. These different experiments are referred to as 'New genus', 'New order' and 'New class' respectively. This allows us to examine the performance in more realistic settings. A drastic drop in the sensitivity and accuracy of the alignment-based methods (MEGAN and PhymmBL) can be seen in the absence of closely related genomes. This is due to the lack of homologous regions, as only 100 kb of sequence were available for the dominant populations. On the other hand, composition-based methods (PhyloPythiaS and Phymm) show better sensitivity and accuracy, of which PhyloPythiaS shows superior performance. This demonstrates the strength of composition-based methods and the ability of PhyloPythiaS to learn accurate models from limited amounts of reference data (Figure 3.1).

3.5.2 SIMULATED SHORT FRAGMENTS DATA SETS

This is one of the most complex tasks in metagenome sample classification; for a real sample corresponding to the task of assigning individual unassembled reads of rare organisms without reference sequences available to correct higher-level clades. The test fragments do not map to any genus in the reference taxonomy (or available reference sequences). The lowest clades that the fragments map to in the reference taxonomy are at varying taxonomic ranks above the rank of genus. Thus, no assignment to a genus-level clade is the optimal result for fragments of this data set; meaning that genus-level assignment specificity can be computed, while sensitivity of assignments, indicates the portion of correctly 'not assigned' test fragments.

The results are summarized in Figure 3.2. As expected, all methods show better performance with increasing fragment length and a trade-off between sensitivity and specificity. Overall, MEGAN shows superior specificity compared to all other methods. MEGAN is conservative due to its LCA algorithm, in the sense that it makes very specific assignments at the cost of sensitivity. Of the sequence composition-based methods, PhyloPythiaS and Phymm, PhyloPythiaS shows better specificity with compromised sensitivity.

Figure 3.1. Average performance for the simMC data set at different taxonomic ranks in four different experiments.

Both Phymm and PhymmBL show rather low sensitivity at the genus level. This is caused by the composition of the test data, for which none of the test fragments belong to any of the genus-level clades that are part of the models. Both methods 'over-bin' by assigning a substantial fraction of sequences to genus-level clades that should rather be left unassigned. It is interesting to note the drastic performance improvement of PhymmBL compared to Phymm for all fragment lengths and at all taxonomic ranks. At family level, which is the lowest taxonomic rank with valid assignments, the improvement in specificity is approximately 12-18% with a bigger effect on shorter fragments, and around 13% improvements in sensitivity. Furthermore, the fact that MEGAN achieves high specificity values indicates that alignment-based sequence similarity information is beneficial for short fragment assignment. For sequence composition, we attribute the degraded performance to the comparatively weak and noisy compositional signal of short fragments.

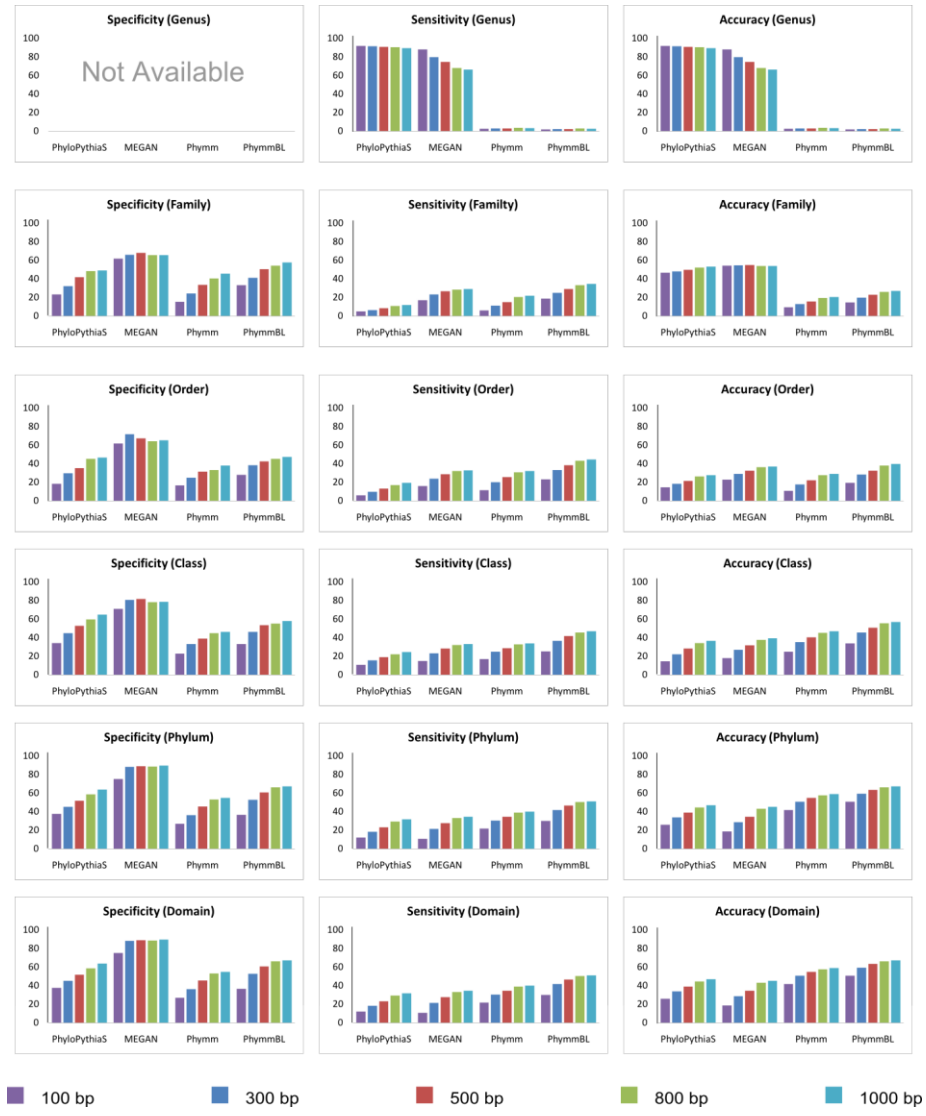


Figure 3.2. Average performance for the simSF data set at different taxonomic ranks.

A “dip” is observed in the specificity at the order level for PhyloPythiaS and other methods. This is due to the construction of the data set. More specifically, the test fragments have varying degree of evolutionary relationship with the reference sequences. This is the reason for non-monotonous behavior of the performance measures over different taxonomic ranks on this data set.

Besides the hold-out experiments described above, we furthermore performed 3-fold cross validation for PhyloPythiaS on the pooled data of complete genome sequences and whole genome assemblies. The data were randomly split into three stratified sets according to their genus affiliations. Genome sequences belonging to one of these sets were used to generate short fragment test data, while the sequences of other two sets were used for training. This procedure was repeated for each of the three sets and assignment accuracy determined. The averaged sensitivity, specificity and accuracy values obtained are reported in Figure 3.3.

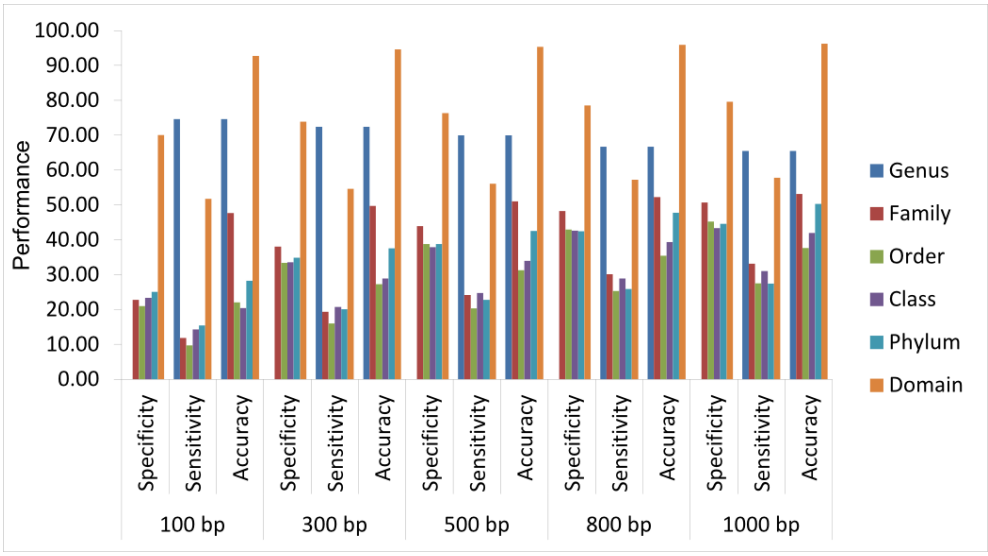


Figure 3.3. Average performance of PhyloPythiaS on the genus-stratified short fragment data sets.

3.5.3 Acid mine drainage metagenome Sample

We compared the PhyloPythiaS generic and sample-specific model assignments with predictions from the NBC web server (http://nbc.ece.drexel.edu/), MEGAN and the best BLASTN hit approach. As MG-RAST and WebCARMA incorporate AMD sequences as reference data, a comparative evaluation by direct submission to these servers would not have ensured strict separation of the reference data and test data. Taxonomic scaffold assignments with PhyloPythiaS and the other tested methods were evaluated based on draft genome assemblies for the five strains and the Fluorescent In-Situ Hybridization (FISH) cell counts published in the original AMD study (Figure 3.4 d, e).

The PhyloPythiaS generic model returned the assignments in less than 5 minutes when accessed via the web server running on a machine with 4 GHz CPU, 4 GB main memory and no competing processes. Most scaffolds were assigned to high taxonomic ranks (taxonomic assignments are shown in Figure 3.4, base-pair accuracy is given in Table 3.2. Taxonomic distance analysis for the AMD metagenome scaffolds assignment.). As with complete scaffolds, bacterial clades were overestimated and archael clades were underestimated (Table 3.2, Supplementary Figure 3). As no reference data were available in model construction for the sample populations, this was expected. Euryarchaeota were identified, but many scaffolds were assigned to phyla Proteobacteria and Verrucomicrobia, instead of to Nitrospirae. The generic model assignments were similar to those of BLASTN in terms of population abundance (Supplementary Figure 5). In contrast, the NBC web server overestimated the abundance of Firmicutes and underestimated that of Euryarchaeota (Figure 3.4 f, Supplementary Figure 6). It might be that the NBC web server performs better on short sequence fragments rather than on longer sequences. In order to check for this possibility, we created fragments of length 500 bp from the AMD scaffolds and obtained their assignments. In this case, the NBC server was accessed in May 2011. The resulting assignments were mapped to the phylum and domain level clades to facilitate visualization (Supplementary Figure 7).

For learning a sample-specific model, we randomly selected ~100 kb of continuous sequences from the five populations as sample-specific training sequences. Specifically, the five strains and corresponding amounts of sample-specific data used were 70 kb for Leptospirillum sp. Group III, 100 kb for Ferroplasma acidarmanus Type I, 100 kb for Leptospirillum sp. Group II '5-way CG', 100 kb for Ferroplasma sp. Type II and 70 kb for Thermoplasmatales archaeon Gpl (G-plasma). Construction of the sample-specific model took slightly less than 7 hours. Assignments with the sample-specific model (Figure 3.4 b, c and Supplementary Figure 4) corroborate well with the taxonomic makeup of this dataset. Both the generic and sample-specific models of PhyloPythiaS produced assignments that were taxonomically consistent and closer to the draft assemblies than those of the BLASTN approach, MEGAN and the NBC server (Figure 3.4, Figure 3.5). Low scaffold consistency for the Leptospirillum sp. Group II '5-way CG' population (0.76) accompanied by low taxonomic distance between correct and predicted taxonomic affiliations (1.73) suggest that there was a certain degree of 'back-and-forth' in assignments between the Leptospirillum clades. In contrast, assignments for the Ferroplasma populations showed high scaffold consistency (>0.95) and higher taxonomic distance between correct and predicted affiliation (>3.7), suggesting that assignments were made to higher ranks (Table 3.2).

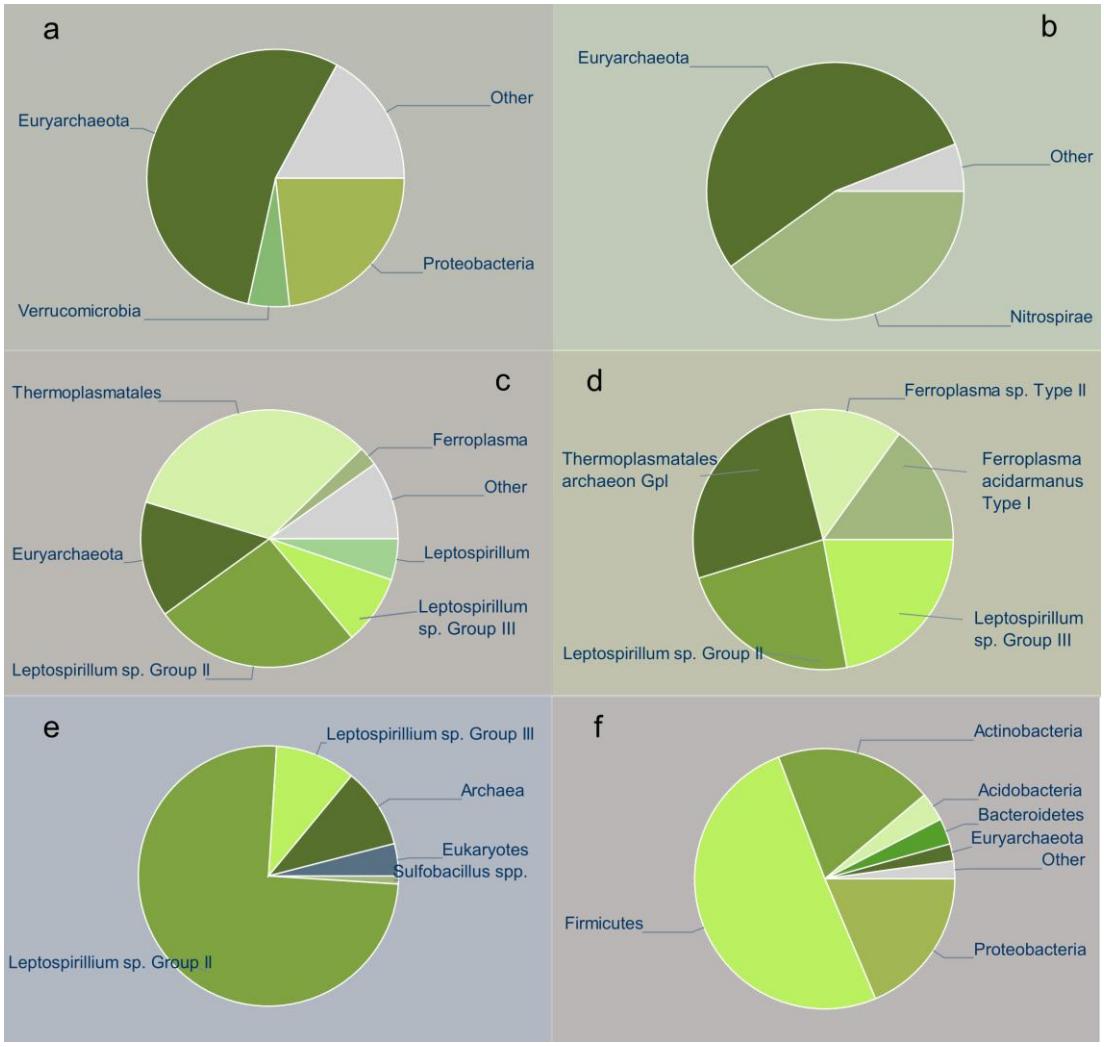


Figure 3.4. Taxonomic assignments of the AMD metagenome scaffolds. Each slice represents number of bases assigned. (a) the PhyloPythiaS generic model at the phylum level, (b) the PhyloPythiaS sample-specific model at the phylum level, (c) the PhyloPythiaS sample-specific model at various ranks, (d) taxonomic reference composition, obtained by alignment of the scaffolds with draft genome assemblies, (e) quantitative cell counts from a FISH study, reproduced from (Tyson et al. 2004) and (f) NBC with N-mer length 15 and Bacteria/Archaea genomes at the phylum level. The "Other" slice represents sequences that were unassigned or assigned at a higher level. Assignments were mapped to phylum level in plots a, b and f for ease of visualization.

Table 3.2. Taxonomic distance analysis for the AMD metagenome scaffolds assignment. The most specific assignments provided by each method were used for this analysis. The correct scaffold assignments (Population)**, were obtained using five strains (three species) whole genome shotgun sequences obtained from NCBI. The methods are PhyloPythiaS sample-specific model** (PPS SS)**, PhyloPythiaS generic model** (PPS G)**,** BLASTN, MEGAN **and naïve Bayesian classifier** (NBC)**. The populations are *Thermoplasmatales archaeon Gpl*** (T)**, *Leptospirillum sp. Group III*** (L1)**, *Leptospirillum sp. Group II '5-way CG'*** (L2)**, *Ferroplasma acidarmanus*** (F1) **and *Ferroplasma sp. Type II*** (F2)**. The numbers in brackets after population name show number of correct scaffolds. The rows signify number of assigned scaffolds** (Assigned)**, the fraction of assignments in the same lineage as the correct taxon** (Const_n_scaff)**, the fraction of base-pairs in the same lineage as the correct taxon** (Const_n_bp) **and average taxonomic distance with respect to draft reference genomes** (Tax Dist)**.**

Method	Measure	Population						
		T (404)	L1 (417)	L2 (126)	F1 (172)	F2 (64)	Micro average	Macro average
PPS SS	Assigned	404	410	118	172	64	--	--
	Const_n_scaff	0.83	0.91	0.76	0.98	0.95	0.89	0.89
	Const_n_bp	0.89	0.94	0.95	0.99	0.99	0.94	0.95
	Tax dist	2.82	1.60	1.73	3.72	3.83	2.11	2.74
PPS G	Assigned	403	414	126	172	64	--	--
	Const_n_scaff	0.81	0.38	0.29	0.97	0.91	0.63	0.67
	Const_n_bp	0.86	0.38	0.11	0.99	0.98	0.62	0.66
	Tax dist	2.96	8.01	7.56	4.46	3.70	4.97	5.34
BLASTN	Assigned	403	416	126	172	64	--	--
	Const_n_scaff	0.13	0.16	0.05	0.07	0.08	0.12	0.10
	Const_n_bp	0.08	0.11	0.01	0.02	0.02	0.05	0.05
	Tax dist	5.65	11.18	11.45	7.97	6.64	7.90	8.58
MEGAN	Assigned	377	306	89	164	63	--	--
	Const_n_scaff	0.38	0.67	0.61	0.24	0.25	0.22	0.43
	Const_n_bp	0.33	0.65	0.57	0.19	0.12	0.37	0.37
	Tax dist	4.16	6.91	6.62	6.98	5.81	3.55	6.09
NBC	Assigned	403	413	126	172	63	--	--
	Const_n_scaff	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Const_n_bp	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Tax dist	11.35	10.97	10.65	14.85	13.63	12.40	12.29

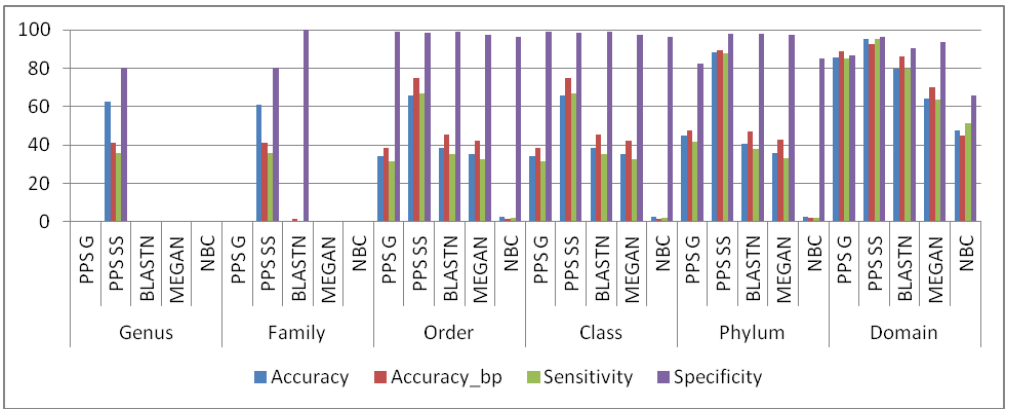


Figure 3.5. Performance of the different methods at six major taxonomic ranks on the AMD metagenome sample. All the methods except PhyloPythiaS in sample-specific mode and BLASTN made only incorrect assignments at genus and family levels. The performance measures are used as defined in section 3.2. The methods compared are the PhyloPythiaS generic model (PPS G), PhyloPythiaS sample-specific model (PPS SS), BLAST best hit (BLASTN), MEGAN and naïve Bayesian classifier (NBC).

3.5.4 TAMMAR WALLABY FOREGUT METAGENOME SAMPLE

For taxonomic sample characterization, sample-specific models were constructed by combining publicly available sequences from NCBI (complete genomes and draft assemblies) with sample-specific data identified based on taxonomic marker genes and sequencing of a scaffold metagenome library. The PhyloPythiaS and PhyloPythia models included a representation for these abundant sample-population in addition to higher-level bacterial and archaeal clades (Supplementary Table 1) (Pope et al. 2010; Patil et al. 2011). Sample-specific data was also incorporated into the training data for PhymmBL and a reference database for BLASTN similarity searches for MEGAN. Note that the PhyloPythia model was built and the assignments were obtained for the (Pope et al. 2010, 2011) studies.

The performance of the different methods for the three abundant populations and the whole sample on average based on the scaffold-contig consistency of the assignments was calculated (Table 3.3). Figure 3.6 and Supplementary Figure 8 depict the scaffold-contig assignment consistency for scaffolds longer than 20 kb for the WG-1 and WG-2 populations for the different methods, respectively. Both PhyloPythiaS and PhyloPythia show a higher consistency than PhymmBL and MEGAN for the three uncultured populations; except that MEGAN has a slightly better consistency for WG-3 (Table 3.3). The overall consistency of MEGAN assignments is higher than for the other methods, but a considerably smaller portion of the sample is characterized (~63%), while the rest remained unassigned. PhymmBL assigned a large portion of the sample (~98%, following PhyloPythiaS ~100%) but shows lower consistency values.

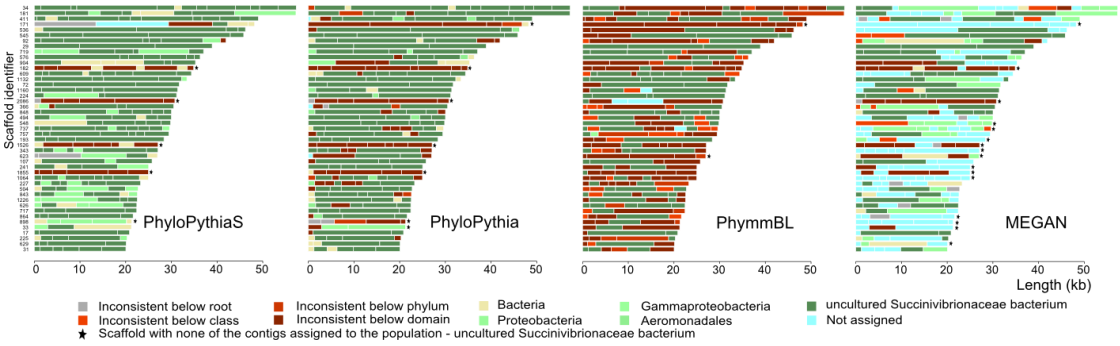


Figure 3.6. Comparison of different taxonomic assignment methods using scaffold-contig consistency for the WG-1 population (uncultured Succinivibrionaceae bacterium) from TW sample. Contig coloring reflects taxonomic assignment consistency with respect to WG-1. Every horizontal bar represents a scaffold and its constituent contigs. Every contig is color coded to represent its consistency with respect to the scaffold assignment. Only scaffolds >=20 kb in length are shown for clarity.

Table 3.3. Performance of different binning methods for the abundant populations in the TW sample. Assignment accuracy is evaluated based on the scaffold-contig consistency. Sample-specific data was used for all methods.

Method	Population	Kilo-bases assigned	Scaffold-contig consistency (% bp)	Scaffold-contig consistency (average taxonomic distance)
PhyloPythiaS	WG-1	2,669.60	97.71	0.38
	WG-2	2,512.93	97.24	0.34
	WG-3	892.65	94.11	0.43
	Total	13,552.86	78.54	0.44
PhyloPythia	WG-1	2,674.70	97.94	0.29
	WG-2	2,326.76	89.75	0.53
	WG-3	870.60	94.70	0.35
	Total	12,830.05	82.90	0.43
PhymmBL	WG-1	3,542.94	69.90	0.72
	WG-2	2,809.81	56.69	1.12
	WG-3	1,005.99	64.59	1.12
	Total	13,286.18	60.78	1.01
MEGAN	WG-1	1,100.20	90.28	0.44
	WG-2	646.19	81.99	0.46
	WG-3	142.69	95.27	0.27
	Total	8,604.92	86.91	0.41

Table 3.4. Effect of sample-specific data on the assignment of the TW sample for PhyloPythiaS and PhymmBL. The "#predictions" columns shows number of predictions obtained using the sample-specific models and for both the sample-specific and the non-sample-specific models. The "#consistent predictions" column shows how many of these predictions are taxonomically consistent with the respective population. The last column shows the average taxonomic distance between the predictions of the sample-specific and non-sample-specific models. For WG-2 PhymmBL without sample-specific data made the specified number of consistent assignments to *Lachnospiraceae* due to relabeled *Ruminococcus*.

Population	Method	#predictions (sample-specific)	#predictions (joint)	#consistent predictions	Average taxonomic distance
WG-1	PhymmBL	530	434	0	8.93
	PhyloPythiaS	477	477	361	5.13
WG-2	PhymmBL	708	690	205	5.37
	PhyloPythiaS	482	482	419	2.05
WG-3	PhymmBL	286	201	0	8.59
	PhyloPythiaS	296	296	266	3.29

PhyloPythiaS and PhyloPythia have comparable consistency. For WG-2, PhyloPythiaS had higher consistency, for WG-1 PhyloPythia performed slightly better. PhymmBL showed lower consistency, both for the dominant populations and the whole sample. PhymmBL generally assigns fragments down to the genus level-clades of the model, which results in lower consistency values.

We evaluated the performance of PhyloPythiaS and PhymmBL in the presence and absence of the sample-specific data. The results indicate PhymmBL's over-binning tendency of assigning most sequences to genus-level clades (Table 3.4). These assignments can be misleading if genera of the dominant sample populations are not included in the reference model. For PhymmBL, out of 530 contigs that were assigned to WG-1, when sample-specific data was included, only 33 contigs were assigned to the consistent parental clade Gammaproteobacteria without sample-specific data, accompanied by a large number of inconsistent assignments in comparison to assignments of the sample-specific model. In contrast, for the same population, PhyloPythiaS assigned 243 out of 477 contigs to the consistent general clade Bacteria, in the absence of sample-specific data, thus avoiding false positive assignments (Supplementary Table 2). Similar observations were made for other populations (data not shown).

In order to investigate difference between the different taxonomic classification methods, we performed a two-tailed Wilcoxon paired sum-ranks tests for different methods on the scaffold-contig consistency and kilo-bases assigned for 230 clades (union of predicted clades by all the methods). The P-values obtained (Table 3.5) show that PhymmBL is significantly different than other methods in both kilo-bases assigned and scaffold-contig consistency. The differences between the methods were visualized using Euler diagrams (Kestler et al. 2008) (Supplementary Figure 9, Supplementary Figure 10).

Table 3.5. Statistical comparison of the assignments of different methods on the TW data set. The bold values indicate pairs where the null hypothesis is rejected at 95% confidence.

Methods	Scaffold-contig consistency	Kilo-bases assigned
PhyloPythiaS – PhyloPythia	**0.0338**	0.4242
PhyloPythiaS – PhymmBL	**5.5454e-09**	**1.7678e-07**
PhyloPythiaS – MEGAN	0.5720	0.8605
PhyloPythia - PhymmBL	**1.1306e-11**	**6.2198e-11**
PhyloPythia – MEGAN	0.0591	0.5781
PhymmBL – MEGAN	**2.0417e-12**	**8.0705e-06**

NUCmer analysis

A representative of WG-1 has been cultured axenically by reverse metagenomics methods, and its genome sequenced (Pope et al. 2011). NUCmer (nucleotide MUMmer) (Delcher et al. 2002) was used by our collaborator Phil Pope to align the contigs predicted as WG-1 by PhyloPythiaS and PhyloPythia, respectively, to the 43 scaffolds obtained for the WG-1 genome (Table 3.6). Overall, 357 of 366 PhyloPythia assignments (98%) align to the reference, with 90.09%, or 1.79 Mbp, of metagenome sequence matching the genome reference. In comparison, 525 of 604 PhyloPythiaS assignments (87%) align to this reference, corresponding to 85.77%, or 1.80 Mbp, of matching sequence. The average percent identity of aligned metagenome contigs with the reference was 98.92% and 98.9%, respectively. The filtered alignment images indicate that the PhyloPythiaS assignments produce a tighter coverage of the reference scaffolds than those of PhyloPythia (data not shown). The most likely reason for this tighter coverage is that PhyloPythiaS assigns many more short contigs than PhyloPythia. However, despite PhyloPythiaS assigning more contigs, a larger fraction of contigs do not align to the reference, and the extra assignments do not significantly increase the overall coverage, as they mostly consist of short contigs. Whilst the reference WG-1 isolate genome is not 100% complete, there is a likelihood of some miss-assignments arising from the additional, shorter contigs that PhyloPythiaS is assigning to WG-1. This is not surprising, given that the accuracy of short contig assignments generally is not comparable to that for longer contigs (see above). Nonetheless, both methods were very accurate in the taxonomic assignment of this population.

Table 3.6. NUCmer analysis of the WG-1 assignments for the TW sample.

Measure	PhyloPythia filtered	PhyloPythia unfiltered	PhyloPythiaS filtered	PhyloPythiaS unfiltered
# contigs aligned	357 (98%)	359 (98%)	525 (87%)	543 (90%)
Length match (bp)	1,798,591	1,941,532	1,803,892	1,972,064
Coverage (%)	90.09	97.28	85.77	93.7
Average IDY (%)	98.92	95.14	98.90	95.50

3.5.5 Human gut metagenome samples

PhyloPythiaS and PhyloPythia models were constructed for 29 (14+15) genus- and family-level clades abundant in the sample and relevant higher-level taxonomic clades (Supplementary Table 3) using data from 5,548 and 3,391 sample-specific contigs and 1,775 microbial complete and draft microbial genomes. For PhyloPythiaS, sample-specific data was selected with active sampling for training, while for PhyloPythia, a subset was taken. PhyloPythia assignments were generated by Alice Carolyn McHardy in a previous study (Turnbaugh et al. 2010). For the training of PhymmBL, only assembled and draft genome sequences were used. Due to

excessive computational requirements of homology searches on this data set, we did not perform assignments with MEGAN.

Contigs from both samples were assigned with PhyloPythiaS, PhyloPythia and PhymmBL and the scaffold-contig assignment consistency was evaluated. PhyloPythiaS and PhyloPythia consistently showed a very similar performance across all taxonomic ranks (Table 3.7). PhymmBL also showed a high scaffold-contig consistency, but, in comparison, lesser amounts of sequence are characterized. This is indeed an interesting result, as no sample-specific data was included for training of PhymmBL. The high consistency observed in the absence of sample-specific training data may be due to the fact that a large number of 122 available gut genome sequences from the relevant taxa are in the public domain and thus could contribute to model quality.

MARKER GENE ANALYSIS

In addition to the analysis of the scaffold-contig consistency, we performed further tests to validate PhyloPythiaS scaffold assignments for the two human gut microbiome samples, relative to the tests of PhyloPythia and control genomes described in (Turnbaugh et al. 2010). These analyses were performed by our collaborator Peter Turnbaugh on the intersection of the scaffolds assigned by both PhyloPythiaS and PhyloPythia. First, the scaffolds assignments were validated based on 30 conserved marker genes with consistent phylogeny to 16S rRNA. All genes from the microbiome bins were assigned to STRING orthologous groups (Jensen et al. 2009). A neighbor-joining tree was built using clustalw (Larkin et al. 2007) version 2.0.12 for each set of marker genes after aligning the translated gene sequences from 122 gut genomes and the binned scaffolds. Individual sequences were assigned to taxa based on the consensus taxonomy of all sequences found at the first node. Additionally, the frequency of consistent taxonomy between database marker genes and nearest neighbor sequences was tallied and used as a control for the frequency of miss-assignment due to alignment errors, improper clustering, and/or disagreement with the marker genes and NCBI taxonomy. Overall the results indicate accurate binning at all evaluated taxonomic levels. PhyloPythiaS showed a high accuracy based on this measure across the ranks from domain to genus for both the TS28 and TS29 samples (Figure 3.7).

Table 3.7. Taxonomic assignments for abundant genera in the human gut metagenome samples. Assignment accuracy is evaluated based on the consistency of taxonomic assignment for contigs of the same scaffold.

Method	Genus-level bin / Population	Kilo-bases assigned		Scaffold-contig consistency (% bp)		Scaffold-contig consistency (average taxonomic distance)	
		TS28	TS29	TS28	TS29	TS28	TS29
PhyloPythiaS	Ruminococcus	13,787.33	13,016.96	95.10	94.68	0.16	0.20
	Faecalibacterium	17,049.71	8,490.69	93.44	90.75	0.18	0.16
	Clostridium	8296.77	3376.53	89.41	95.74	0.24	0.22
	Eubacterium	8840.37	2515.17	98.05	76.63	0.10	0.30
	Dorea	2,443.36	1,323.47	98.75	96.05	0.11	0.30
	Bifidobacterium	4,948.32	4,760.12	98.51	99.97	0.08	0.05
PhyloPythia	Ruminococcus	16,879.06	14,918.45	94.78	90.18	0.15	0.29
	Faecalibacterium	19,962.39	9,372.68	94.80	85.72	0.28	0.25
	Clostridium	11,797.44	4,097.59	77.42	85.62	0.39	0.45
	Eubacterium	10,138.96	1,859.18	97.12	89.78	0.16	0.51
	Dorea	3,412.84	1,511.66	97.21	82.30	0.11	0.49
	Bifidobacterium	4,946.77	4,767.18	98.40	99.78	0.06	0.03
PhymmBL	Ruminococcus	6,613.42	5,694.06	96.11	94.87	0.10	0.09
	Faecalibacterium	15,302.09	6,423.28	94.09	93.96	0.12	0.07
	Clostridium	13,246.25	4,917.47	87.30	92.22	0.22	0.19
	Eubacterium	5,624.48	1,337.88	98.01	85.77	0.08	0.26
	Dorea	3,118.58	1,381.38	97.61	82.95	0.05	0.21
	Bifidobacterium	5,057.49	4,757.60	97.96	99.93	0.11	0.03

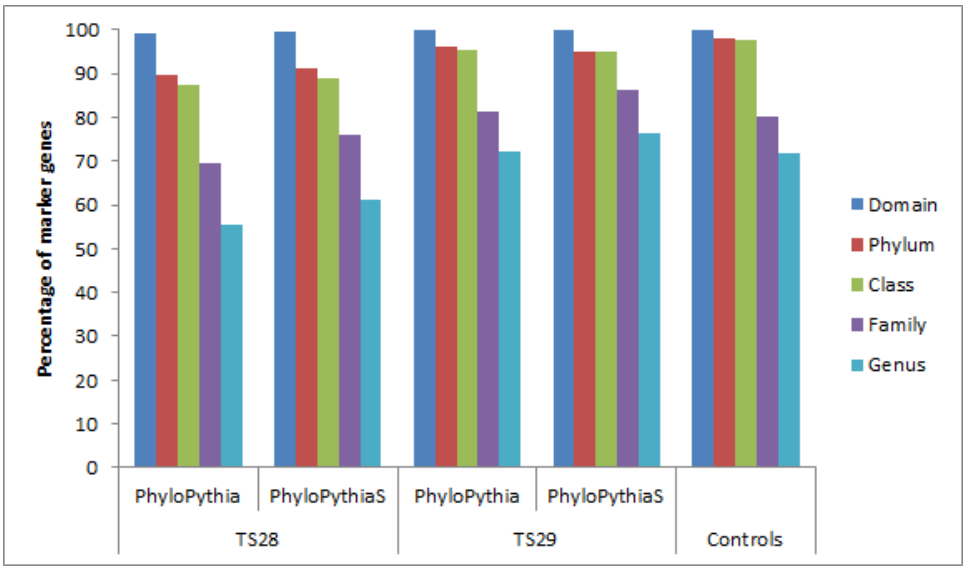


Figure 3.7.Marker gene validation for the human gut metagenome sample assignments.

CD-HIT ANALYSIS

Peter Turnbaugh furthermore used the CD-HIT (Cameron, Bernstein & Williams 2007) to cluster the protein sequences of the gut samples and 122 gut genomes at 60% identity. The taxonomic consistency of genes within these clusters and the respective bin assignments was then analyzed. Both PhyloPythiaS and PhyloPythia showed a high consistency of taxonomic bin assignments within protein clusters (Figure 3.8).

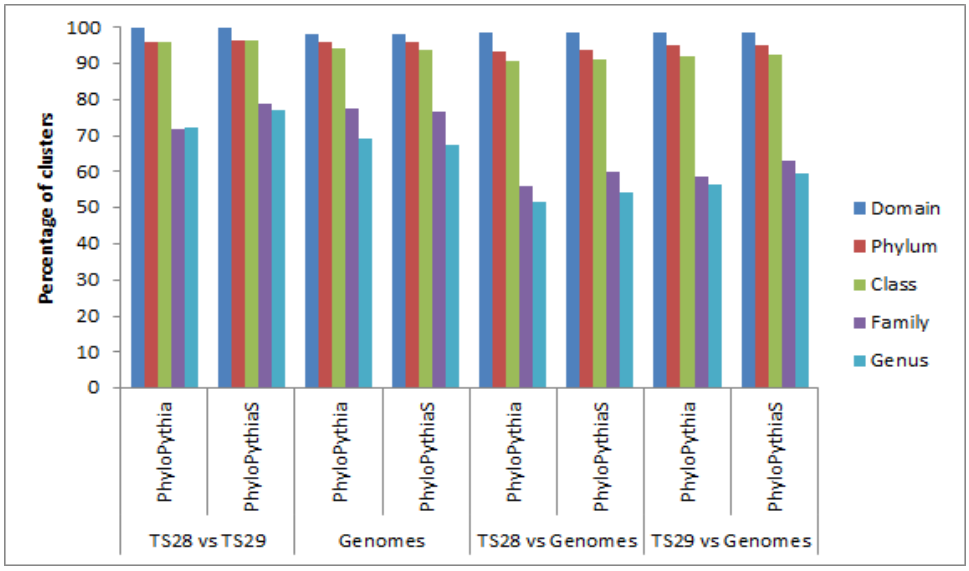


Figure 3.8. Validation for the human gut metagenome sample assignments using CD-HIT (fraction matched).

3.5.6 COW RUMEN METAGENOME SAMPLE

The scaffolds from the CR sample were taxonomically assigned using the generic mode as a multiplex sample (section 2.5) and the combined predictions were visualized. The majority of the scaffolds were assigned to the orders Bacteroidales, Clostridiales, Bacillales, Spirochaetales, Methanomicrobiales, Methanosarcinales, Sulfolobales, Selenomonadales and Rhizobiales (Figure 3.9). We measured the assignment consistency as the number of base-pairs of these scaffolds consistently assigned by the generic model to the order-level clades of the respective genome bins. Taxonomic distances of the predictions were calculated relative to the reported orders for the genome bins (Table 3.8). Overall the generic model made consistent assignments for the majority of scaffolds. In particular, this was the case for genome bins of order-level clades with substantial numbers of reference genomes available, while assignment consistency was lower for clades covered by fewer reference genomes. Seven of the 15 bins were more than 90% consistent, four of them even to 100%. Five bins showed low consistency. In particular, we observed that the Clostridiales and Myxococcales genome bins were less consistent than bins of the other three orders. For Myxococcales this is likely because fewer sequenced genomes were available for training of the generic model (given the number of species with sequenced genomes for all five clades). For the Clostridiales, this might be due to genomic differences of the species represented by the genome bins to the sequenced Clostridiales genomes used as reference (mean GC content of 50% versus a mean GC content of 36%). However, regardless of the exact nature of the assigned taxonomic affiliation,

scaffolds of a particular bin tended to be homogeneously assigned to the same clade by the generic model, varying from 44% to 100% of the scaffolds for the different bins. The predictive accuracy of the overall assignment can likely be further improved by construction of a sample-specific model, as we showed for AMD, TW and HG samples.

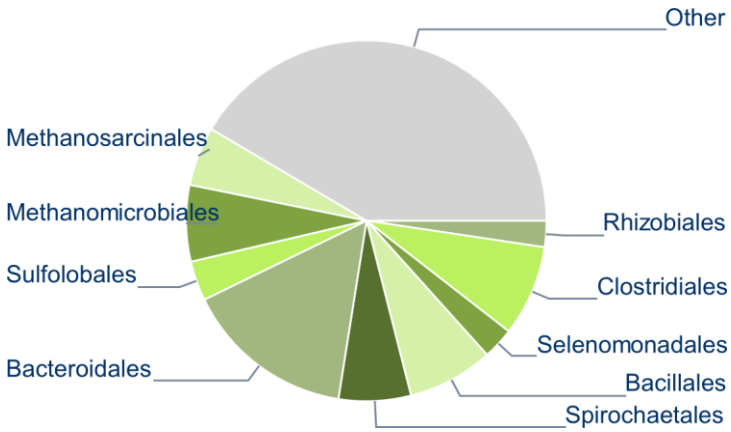


Figure 3.9. Taxonomic assignments of the cow rumen metagenome scaffolds with the PhyloPythiaS generic model. This data-set contained 26,042 scaffolds in total. The assignments are shown at the order level. Each slice represents the total number of bases assigned to an order. The "Other" slice represents sequences that were either assigned at a higher level or were unassigned.

3.6 EXECUTION TIME ANALYSIS

Empirical analysis of execution times, performed on a machine with 4 GHz CPU, 4 GB main memory and no competing processes, determined that PhyloPythiaS requires 0.08-0.1 seconds for the assignment of 0.1-10 kb fragments (Figure 3.10). This corresponds to a 3- to 46-fold and 5- to 68-fold improvement in comparison to MEGAN and PhymmBL, respectively. For characterization of a 13 MB assembled metagenome sample, PhyloPythiaS showed 22-fold, 85-fold and 106-fold speed increase in comparison to PhyloPythia, MEGAN and PhymmBL, respectively (Table 3.9). The efficiency of PhyloPythiaS at the test time is due to the linear nature of the inference that only requires computing the dot product between the input example and the learned weight vector in the joint feature space. As PhyloPythiaS models require only a subsample of the reference data for accurate assignment, in the future, training times will not necessarily be substantially impacted by increases of sequence data, contrary to alignment-based approaches.

Table 3.8. Taxonomic distance and consistency analysis of the 15 genome bins from the cow rumen metagenome consisting of 466 scaffolds in total. The first three columns describe the dataset while the last three columns summarize the predictions of the PhyloPythiaS generic model. The last three columns show the average taxonomic distances between the predicted order and the correct order (Tax Dist), the consistency calculated based on the fraction of assigned scaffolds (Const_n_scaff) and the consistency calculated based on the fraction of assigned base-pairs (Const_n_bp). See 'Results' for the definitions of taxonomic distance and consistency. The micro average is the average value over all scaffolds and the macro average represents the average over the genome bins.

Genome bin	Correct order	#Scaff	PhyloPythiaS generic model prediction		
			Tax Dist	Const_n_scaff	Const_n_bp
AC2a	Bacteroidales	20.000	0.000	1.000	1.000
AJ	Bacteroidales	22.000	0.000	1.000	1.000
AMa	Spirochaetales	19.000	0.000	1.000	1.000
AQ	Bacteroidales	24.000	0.000	1.000	1.000
AH	Bacteroidales	26.000	0.231	0.962	0.990
ATa	Clostridiales	32.000	0.625	0.906	0.967
AGa	Bacteroidales	35.000	0.743	0.886	0.938
BOa	Clostridiales	42.000	1.738	0.690	0.776
AFa	Spirochaetales	28.000	1.893	0.714	0.759
APb	Clostridiales	55.000	3.636	0.382	0.454
AS1a	Clostridiales	53.000	5.245	0.189	0.114
AIa	Clostridiales	22.000	6.682	0.182	0.086
ADa	Myxococcales	20.000	3.100	0.250	0.076
AN	Clostridiales	27.000	3.704	0.074	0.046
AWa	Clostridiales	41.000	7.073	0.000	0.000
Macro average	--	31.067	2.311	0.616	0.614
Micro average	--	--	2.693	0.560	0.613

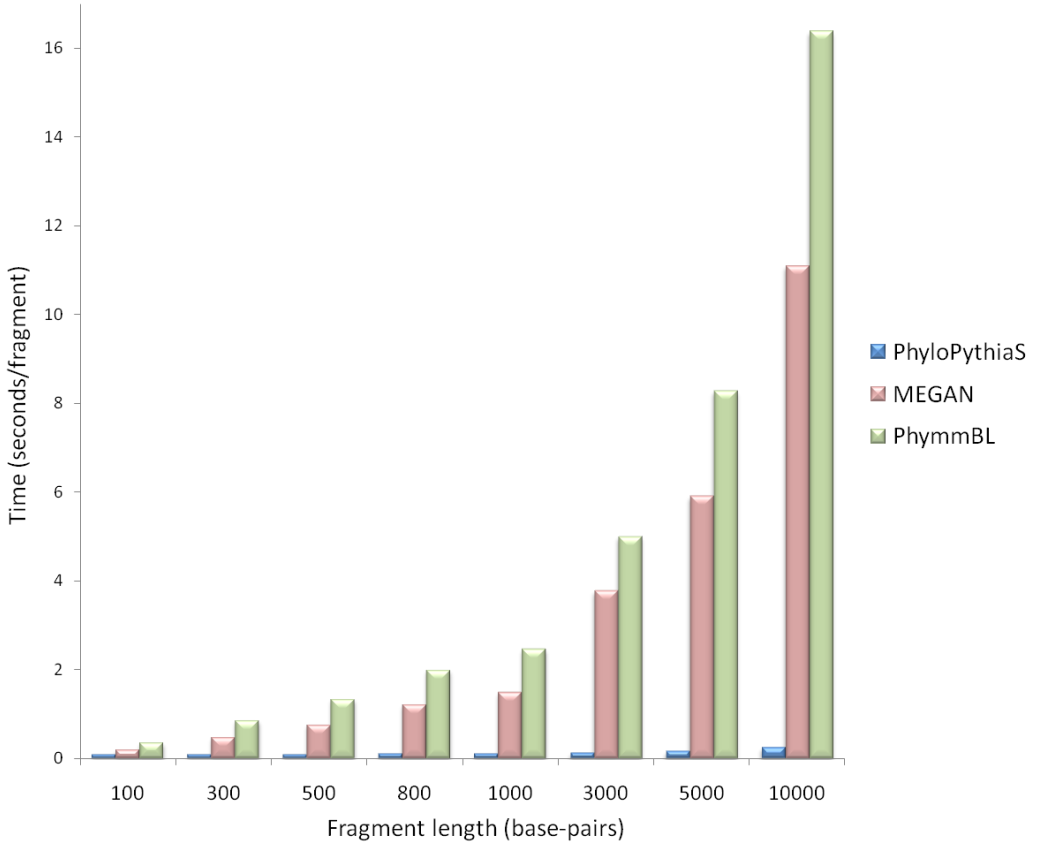


Figure 3.10. Empirical execution time evaluated on a Linux machine with 3 GHz processor and 4 GB main memory. Results for MEGAN and PhymmBL were determined with a reference database of size 2.1 GB.

Table 3.9. Execution time comparison for different methods for characterization of the three real metagenome samples. The sample sizes are approximately 16 Mb, 113 Mb and 72 Mb for TW, TS28 and TS29 respectively.

Method	Time (DD:HH:MM:SS)		
	TW	TS28	TS29
PhyloPythiaS	00:00:08:36	00:01:13:43	00:00:46:28
PhyloPythia	00:03:12:43	01:08:04:25	00:21:18:27
PhymmBL	00:15:09:51	07:13:54:01	04:15:53:44
MEGAN	00:12:10:14	--	--

3.7 CONCLUSIONS

Some general conclusions can be drawn from the experiments performed on the simulated and real metagenome data sets with regards to the closeness of the reference data to the sequences in the sample and sequencing technology used.

When closely related complete genomes sequences are available for the populations in the metagenome sample, alignment-based methods are at an advantage as the sample fragments can be aligned to the respective reference genomes with high confidence. This was observed for the simMC data set in the 'known species' experiment, where complete genome sequences from NCBI were used as reference data for model training with exception of the genomes used to create the simMC data set. Though the exact genomes were removed, the reference data included genomes of either same species (for *Rhodopseudomonas palustris* and *Xylella fastidiosa*) or same genus (for *Bradyrhizobium sp. BTAi1*). At lower taxonomic ranks (genus and family) alignment-based and hybrid methods showed higher sensitivity and accuracy compared to the composition-based methods. At higher taxonomic ranks the sensitivity and the accuracy of all methods became more similar. PhyloPythiaS maintained high specificity at all taxonomic ranks, while other methods except PhyloPythia generally showed lower specificity at lower taxonomic ranks. Similarly, for the two human gut metagenomes the high scaffold-contig consistency obtained by PhymmBL without sample-specific sequences is likely due to the large number of gut genome sequences from related taxa (122 in total) available as reference.

In the taxonomic assignment task for metagenomic data it is more realistic to consider that complete genome sequences of the dominant populations are not available as reference as most of the microorganism diversity is still unknown. Therefore, often only distantly related genomes are available and in some cases it is possible to obtain limited amounts of sample-specific data for the dominant populations by phylogenetic analysis of conserved marker-genes for the sample or sequencing of additional fosmid libraries. We simulated three such scenarios using the simMC data set; 'New genus', 'New order' and 'New class', by retaining 100 kb randomly selected contiguous fragments for dominant populations and removing all reference genomes at the corresponding ranks. In the absence of the closely related genomes the alignment-based and hybrid methods showed a drastic drop in the sensitivity and accuracy. On the other hand, composition-based methods showed better sensitivity and accuracy. This demonstrates strength of composition-based methods and the ability of PhyloPythiaS to learn accurate models from limited amounts of reference data. When no closely related or sample-specific data is available PhyloPythiaS tends to make assignments at higher taxonomic ranks. This is a desired behavior as assignments to lower ranks can be misleading in these cases. This suggests that PhyloPythiaS is better at assigning fragments of the 'known unknowns' in metagenome data sets and is robust with respect to the reference data.

Furthermore, with many high-throughput sequencing technologies being developed, we also evaluated whether PhyloPythiaS copes with the different technology-specific errors and read lengths. The technologies produce reads of different lengths and qualities, potentially affecting performance of taxonomic assignment methods. We tested sequences generated with three

technologies; Sanger, 454/Roche and Illumina, and found that regardless of the technology used all samples were characterized consistently. We expect PhyloPythiaS to work equally well with assembled sequence data from other technologies with similar sequencing error rates, such as the SOLiD (Applied Biosystems) platform (Valouev et al. 2008). It should be noted that the performance of PhyloPythiaS on sequence fragments with high error rates is still unexplored. Although it is possible to perform assignments for short sequences (<1000 bp), like with other methods, these assignments are less accurate than those for longer sequences and often to higher ranking taxa only. Therefore, we advise that short reads should be assembled into longer contigs before performing assignments with PhyloPythiaS.

4 GENOME TREE INFERENCE

Understanding the evolutionary relationships between organisms is vital for their in-depth study. Gene-based methods are often used to infer such relationships, which are not without drawbacks. One can now attempt to use genome-scale information, because of the ever increasing number of genomes available. This opportunity also presents a challenge in terms of computational efficiency. Two fundamentally different methods are often employed for sequence comparisons, namely alignment-based and alignment-free methods. We used genome-scale sequence information to infer taxonomic distances between organisms without additional information such as gene annotations. We propose a method to improve genome tree inference by learning specific distance metrics over the genome signature for groups of organisms with similar phylogenetic, genomic or ecological properties. Specifically, our method learns a Mahalanobis metric for a set of genomes and a reference taxonomy to guide the learning process. By applying this method to more than a thousand prokaryotic genomes, we show that, indeed, better distance metrics could be learned for most of the 18 groups of organisms tested here. Once a group-specific metric is available, it can be used to estimate the taxonomic distances for other sequenced organisms from the group. This study also presents a large scale comparison between ten methods - nine alignment-free and one alignment-based.

4.1 INTRODUCTION

In this chapter we address the problem of inferring distances between whole genome (genic + nongenic) sequences to recover their evolutionary relationships in the form of a tree that we will refer to as the genome tree. The evolutionary relationships between different organisms, and hence their genomes, are typically represented in the form of a phylogenetic tree. Phylogenies are often inferred from individual gene sequences, such as the highly conserved small subunit ribosomal RNA (Woese and Fox 1977) or from a set of conserved orthologous genes (Ciccarelli et al. 2006; Wu and Eisen 2008). Phylogenies inferred from different genes or gene sets often disagree with each other and only show a plausible evolutionary history for the genes used which is not necessarily the evolutionary history of the analyzed taxa (Hasegawa and Hashimoto 1993; Karlin and Cardon 1994). Furthermore, to apply gene-based methods, one must first identify orthologous genes from different organisms which can be difficult due to evolutionary processes such as gene loss, duplication and horizontal transfer (Doolittle 1999). With the availability of a large number of completely sequenced genomes whole genome based methods were proposed to alleviate the shortcomings of gene based methods. Various properties of the genome such as gene content, gene order, whole genome sequence similarity and nucleotide composition biases have been used to measure distances between genomes, see (Coenye et al. 2005; Delsuc, Brinkmann, and Philippe 2005; Snel, Huynen, and Dutilh 2005) for recent reviews. In this work we focused on the analysis of sequence based methods for which no additional information, such as gene annotations, is required.

In section 1.4.2 we described the concept of the genome signature and its advantages over alignment-based comparison. However, the strength of the phylogenetic signal provided by the genome signature varies for different groups of genomes (Mrazek 2009). An important property of the genome signature is that it allows comparison between non-homologous

sequences. For a given species or higher-level clade, it allows an accurate distinction for 1000 bp or longer segments, with longer segments encoding a stronger signal (see Chapter 3) (Deschavanne et al. 1999; Sandberg et al. 2001; Jernigan and Baran 2002; McHardy and Rigoutsos 2007; Patil et al. 2011).

As more whole genome sequences are deposited in public databases, in comparison of alignment-based approaches, the computationally less expensive alignment-free methods become increasingly attractive for the analysis of large-scale data sets (Höhl, Rigoutsos, and Ragan 2006; Yang and Zhang 2008). Some limitations of the genome signature in this context have been pointed out, such as a lower correlation with phylogenetic distance, especially for distantly related genomes (Mrazek 2009), as well as the clustering of distantly related genomes with similar GC-content (Takahashi, Kryukov, and Saitou 2009) (see (Coenye and Vandamme 2003; Pride et al. 2003; van Passel et al. 2006)).

In alignment-free sequence comparison, most research has focused on the identification of the appropriate length for oligonucleotides (Karlin and Burge 1995; Karlin, Mrazek, and Campbell 1997; Kirzhner et al. 2002; Pride et al. 2003; Wu, Huang, and Li 2005; Mrazek 2009; Sims et al. 2009; Takahashi, Kryukov, and Saitou 2009), normalization procedures (Hao and Qi 2003; Xu and Hao 2009) and different distance functions (Wu, Burke, and Davison 1997; Kirzhner et al. 2002; Höhl, Rigoutsos, and Ragan 2006). The genome signature is inherently redundant due to the reverse complementarity of the DNA strands. Under the influence of selection, all oligonucleotides might not be equally important in taxonomic distance calculation, in case they evolve at different rates. These issues have not been given enough attention. Based on the hypothesis that a group of genomes with similar phylogenetic, genomic or ecological attributes might have specific oligonucleotide weights that reflect their importance in distance calculation, we propose a novel method that aims at improving genome signature-based inference of genome trees. Thus, our goal is to enhance the signal for a group by learning group-specific oligonucleotide weights. We propose a supervised distance metric learning method that exploits the structure of known reference taxonomy to guide the learning process (see Materials and Methods). We use the taxonomy as reference for calculation of phenetic distances, rather than a phylogeny (such as one inferred from the 16S rRNA gene), due to its "polyphasic" nature that takes genotypic and phenotypic aspects into account (Vandamme et al. 1996) and not to bias our analysis towards possible shortcomings of gene based methods. However, we verified that phenetic distance strongly correlates with phylogenetic distance (see Materials and Methods).

The aim of our method is to identify a diagonal positive semi-definite matrix parameterizing the Mahalanobis distance metric such that it maximizes the Spearman's rank correlation coefficient between the resulting distances and the phenetic distances within the reference taxonomy. The phenetic distances were calculated similarly to the path loss defined in section 2.3.2 (see Eq. 2.13). The distance metric learning problem is posed as a regularized optimization problem (see section 4.2.6 below). We defined 18 groups based on phylogenetic, genomic or ecological factors. Contrary to other genome tree inference methods, our aim is to improve the performance for a group of genomes defined by a common factor, such as genome-wide GC-content or habitat, and not to reconstruct the entire tree of life. When the species composition or ecological characteristics of the organisms at hand is approximately

known, one can learn a group-specific distance metric using other available reference data. Once a specific distance metric has been learned, it can be employed for the analysis of novel genome sequences from the same group.

Various methods have been proposed for the evolutionary comparisons of entire genomes or large genome segments, including alignment-free methods (Burge, Campbell, and Karlin 1992; Karlin and Cardon 1994; Kirzhner et al. 2002; Pride et al. 2003; Qi, Wang, and Hao 2004; Sims et al. 2009; Takahashi, Kryukov, and Saitou 2009; Li, Xu, and Hao 2010) and the alignment-based methods, such as the genome blast distance phylogeny (GBDP) (Henz et al. 2005). A direct comparison between genome tree inference methods is lacking, especially with the alignment-based method GBDP. Therefore, in addition to proposing a new method, we also present a large scale numerical comparison of the performance of ten genome tree inference methods, including nine alignment-free methods and one alignment-based method.

4.2 MATERIALS AND METHODS

Continuing the notation used in section 1.4.2 each genomic signature is denoted with a pattern *lknm*, where *k* is the oligonucleotide length and *m* is the length of oligonucleotides used for normalization. Thus, for example, the tetranucleotide signature normalized using base frequencies is denoted as l4n1. The notation is optionally followed by the alphabet used (e.g. “ry”) if an alphabet other than nucleotide was used.

We used 1076 complete microbial genome sequences available from NCBI in April 2010 for this study. This corresponds to 578,350 pairs of taxa to compare in terms of their taxonomic and genomic distances. To compute pair-wise distances between species, nine alignment-free methods for computing pair-wise genome distances were tested; the Euclidean distance based on the l4n1 genome signature, the Euclidean distance based on the l4n1 signature after dimensionality reduction with PCA, the Euclidean distance based on the l6n1 signature, CVTree with the l6n5,4 signature (Hao and Qi 2003), the compositional spectrum based on the l10r2 signature and n=200 (Kirzhner et al. 2002) and the feature frequency profile based on the RY alphabet with l=10 (Sims et al. 2009). In addition we also evaluated the GBDP method based on BLAST alignments (Henz et al. 2005), for which we aligned all pairs of genomes. Pair-wise alignments between the nucleotide sequences were generated with the “bl2seq” program (version 2.2.18) with default parameters. Details on these methods are provided in section 4.2.10.

The genomes were subsequently classified into 18 groups according to the following five factors: Phylum membership (4 groups), genomic GC-content (3 groups), habitat (5 groups), temperature range (3 groups) and oxygen requirement (3 groups). For each of these factors, the groups were exclusive (Supplementary Table 4).

4.2.1 GENOMES, TAXONOMY AND ECOLOGICAL INFORMATION

Genome sequences were obtained from GenBank (http://www.ncbi.nlm.nih.gov/genome). The taxonomy from the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/) and the ecological information was obtained with the NCBI *lproks* service (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) (Sayers et al. 2009).

4.2.2 GENOME SIGNATURE

The genome signature represents a sequence as a point in a multi-dimensional metric space. The dimensionality of the space is defined by the size of the alphabet and the length of oligonucleotides. In our case the alphabet comprises four nucleotides (A, T, G and C) and the oligonucleotide length considered is four, which gives rise to a 4^4 dimensional space. The vector representation of sequences allows application of distance metric functions to these points to uncover their interrelationships. We used the tetranucleotide signature vector normalized based on mononucleotide frequencies (l4n1) for learning group-specific metrics. The elements of this signature for a sequence N are defined in Eq. 2.17, which is repeated below for convenience;

$$\rho^{*l4n1}_{abcd|\mathrm{N}} = \frac{\mathrm{fr}^*(abcd)}{\mathrm{fr}^*(a)\,\mathrm{fr}^*(b)\,\mathrm{fr}^*(c)\,\mathrm{fr}^*(d)}$$

As before, here fr* denotes frequency of the oligonucleotides averaged over both strands.

4.2.3 PHENETIC DISTANCES BETWEEN PAIRS OF TAXA IN THE REFERENCE TAXONOMY

As our target variable, or reference distance, we used the phenetic distance between taxa in the NCBI taxonomy. The phenetic distance between a pair of taxa was defined as the maximum number of edges in the path between one of the taxa in the pair and their lowest common ancestor. Seven major taxonomic ranks; species, genus, family, order, class, phylum and superkingdom, were used to calculate the phenetic distances. Note that the number of edges to the lowest common ancestor can differ in the NCBI Taxonomy for two taxa at a given rank, due to missing internal nodes on the path from these taxa to their lowest common ancestor. The matrix containing pair-wise phenetic distances will be denoted as $\mathbf{D}_{\mathbf{TAX}}$.

To compare the phenetic distances with phylogenetic distances, aligned 16S rRNA gene sequences were obtained from the greengenes database (http://greengenes.lbl.gov) (DeSantis et al. 2006). When multiple genes were available for an organism only the first was chosen. In total, genes for 887 organisms were identified. Pair-wise distances between the aligned genes were calculated with the "DNADIST" program in the Mothur package (Schloss et al. 2009). The phenetic distances showed a strong correlation with the phylogenetic distances (Pearson's R= 0.84 and Spearman's ρ=0.81, P=0.001 based on 999 permutations). This suggests that our results should be valid if 16S rRNA distances were used instead of phenetic distances.

4.2.4 COMPARING TREES BASED ON COPHENETIC CORRELATION

The correlation between two tree path metrics has been used to compare tree topologies (Pazos and Valencia 2001; Kuramae et al. 2007). We here used a similar approach to search for a distance metric which best approximates the phenetic distances between pairs of taxa in a given reference tree. As we were interested in the topology of the trees and not branch lengths, we used Spearman's rank correlation coefficient to quantify the agreement between the phenetic distances in the reference topology and pair-wise distances between genome

sequences. Although commonly used, Pearson correlation between distance matrices does not always imply better topology recovery (Lapointe and Legendre 1992). Spearman's rank correlation is furthermore more appropriate when outliers are present and there is a non-linear relationship between the variables. As we are calculating correlation between two symmetric matrices, they are first vectorized using either the upper or lower half triangle. Spearman's ρ is calculated on the ranks x_i and y_i of elements in the vectorized distance matrices according to;

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad \textbf{Eq. 4.1}$$

The correlation between a data-derived matrix of pair-wise distances and a phenetic distance matrix is also known as the cophenetic correlation coefficient (CPCC) (Sokal and Rohlf 1962). The CPCC has been used for assessing how well tree topologies inferred with different hierarchical clustering methods agree with a matrix of pair-wise distances inferred from the data. Here we use it to evaluate how well different data-derived distance metrics agree with phenetic distances between pairs of taxa in reference taxonomy. Although typically Pearson correlation is used to calculate CPCC, the use of rank based correlation has been proposed before (Johnson 1967; Mrazek 2009).

4.2.5 Topological distance between trees

As the cophenetic correlation might not directly correspond to topological similarity (Farris 1969) we also calculated topological distances between trees. The topological distances between trees were calculated using the normalized quartet distance, as implemented in the program QDist (Nielsen et al. 2011) version 2.0, downloaded from http://birc.au.dk/software/qdist/.

Note that an increase in congruence between tree topologies results in an increase in the cophenetic correlation coefficient and a decrease in the quartet distance. The cophenetic correlation was used also as the optimization criterion as described in the following section.

4.2.6 Distance metric learning

The Euclidean distance metric is often used to calculate dissimilarities for data that can be represented as points in a multi-dimensional metric space. However, it may not be ideal to infer taxonomic distance between pairs of genomic signatures. This is particularly true when some of the variables are more important than others or when some dimensions are correlated and/or have different scales, for instance, some different genomic features could be subject to different evolutionary constraints and evolve at different rates. In such cases, a more suitable distance metric than the Euclidean metric can be learned from data. Originally, distance metric learning was proposed for clustering applications where *side information* such as similarity and dissimilarity constraints is available (Xing et al. 2002). The information available in our case is the phenetic distances between pairs of taxa in the reference taxonomy.

Distance metric learning can be viewed as a transformation of the input space into another (possibly lower dimensional) space, in which the Euclidean distance between the points represent as accurately as possible the target relationships. Practically, this can be achieved by using the Mahalanobis distance function. The Mahalanobis distance is a distance metric, parametrized by a positive semi-definite matrix $\mathbf{M}$. The Mahalanobis distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined in Eq. 1.18 and is repeated below for convenience;

$$\mathrm{Mahal}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{M}(\mathbf{x}-\mathbf{y})}$$

We propose learning a diagonal matrix $\mathbf{M}$ with nonnegative entries that maximizes the performance criterion; that is the Spearman's correlation coefficient between the resulting $n\times(n\text{-}1)/2$ pair-wise Mahalanobis distances for n analyzed genomic signatures with the corresponding target phenetic distances. The entries in the target distance matrix, $\mathbf{D_{TAX}}$, were defined as described above. The diagonal elements of the matrix $\mathbf{M}$ represent the relative weights for the corresponding oligonucleotides. The Euclidean distance is a special case of the Mahalanobis distance, when it is parameterized by an identity matrix and the Mahalanobis distance corresponds to a weighted Euclidean distance, when it is parameterized with a diagonal matrix. Let us define a function $\mathrm{d_{Mahal}}$ which returns all pair-wise Mahalanobis distances between a set of vectors $\mathbf{S}$ given a parameterizing matrix $\mathbf{M}$.

Even though a learned metric works well for a given set of signatures (training data) it might not provide improvement for novel signatures (test data). Such over-fitting is not desirable and hence we pose the learning problem as a regularized optimization problem;

Optimization problem $Metric$: Given a training set $\mathbf{S}=\{(\mathbf{x}_i)\}_{i=1}^{n}$ $\mathbf{x}_i\in\mathbb{R}^p$, $\mathbf{D_{TAX}}\in\mathbb{R}^{n\times n}$ and $\lambda\geq 0$

$$\min_{\mathbf{M}}(1-\rho(\mathrm{d_{Mahal}}(\mathbf{S},\mathbf{M}),\mathbf{D_{TAX}}))+\lambda\frac{\sum_{i=1}^{p}M_{ii}}{p} \qquad \textbf{Eq. 4.2}$$
$$s.t.\quad 0\leq M_{ii}\leq 1 \quad i\in 1\ldots p$$

Here p is the number of oligonucleotides and $\mathbf{S}$ is a matrix with each row representing a genomic signature. While first term in the objective function maximizes correlation, the second term is a regularizer that controls complexity of the solutions in terms of the L1-norm of the diagonal entries of M. Thus higher values of λ (λ>=0) will lead to sparse diagonal entries. As only the relative contributions of the oligonucleotides and not their absolute magnitudes are important, the diagonal entries of $\mathbf{M}$ were constrained to values within the interval [0, 1], to allow comparisons between solutions for different experiments. The parameter λ was varied in the set {0, 0.1, 1, 10}. For each value in the grid, a 3-fold cross-validation procedure was performed on randomly partitioned training data as follows; three metrics were learned separately by excluding each of the three partitions and the generalization performance was assessed with the Spearman's correlation between the target distances and the distances with the learned metric on the excluded partition. The resulting three correlations for each λ value were averaged to get an estimate of the generalization performance. The value with the

highest generalization performance was chosen to learn a metric on the complete training data. The aim of the regularizer here is obtaining generalizable solutions and not to enforce sparse solutions. Thus, if a less sparse solution yields a higher generalization performance (as estimated by cross-validation) than a more sparse solution, then the less sparse solution is selected. Note that although it is possible to formulate the optimization problem we describe here with a weight vector instead of the matrix $\mathbf{M}$, the more general formulation clarifies that this method is easily adaptable for learning a full matrix.

We used the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen et al. 2003) as the optimization procedure, though any other global optimization procedure can be used. As this optimization problem is non-linear and non-convex, gradient-based optimization techniques are not appropriate. The python code for CMA-ES was obtained from the website http://www.lri.fr/~hansen/cmaes_inmatlab.html. The tolerance for solution improvement was set to 1e-3 and the number of iterations was set to 500 during cross-validation and 1000 for learning the metric with a selected λ. Only the diagonal of the covariance matrix was adapted to reduce the computational complexity. The population size for CMA-ES was set to 20 and the step-size to 0.5.

4.2.7 SIGNIFICANCE TEST FOR CHANGE IN CORRELATION

The significance of change of the correlation coefficient was assessed with the Hotelling-Williams test between dependent variables (Steiger 1980). Specifically, we tested whether the CPCC of one metric was significantly different from the CPCC of another metric.

4.2.8 MEASURES OF GROUP PHYLOGENETIC STRUCTURE (NRI AND NTI)

We calculated two metrics of group phylogenetic structure. The metrics; net relatedness index (NRI) and nearest taxon index (NTI) correspondingly quantify the distribution of the taxa relative to a phylogeny (Webb et al. 2002). They were calculated as follows;

$$\mathrm{NRI} = -1 \times \frac{\mathrm{mean}(\mathbf{a}_{\mathbf{obs}}) - \mathrm{mean}(\mathbf{a}_{\mathbf{n}})}{\mathrm{sdev}(\mathbf{a}_{\mathbf{n}})} \qquad \textbf{Eq. 4.3}$$

$$\mathrm{NTI} = -1 \times \frac{\mathrm{mean}(\mathbf{b}_{\mathbf{obs}}) - \mathrm{mean}(\mathbf{b}_{\mathbf{n}})}{\mathrm{sdev}(\mathbf{b}_{\mathbf{n}})} \qquad \textbf{Eq. 4.4}$$

Here **a** is a vector containing distances between all pair-wise taxa and **b** is a vector containing distances between all taxa to their nearest taxon, with the same characteristic. The suffix obs denotes observed distances and the suffix n denotes expected distances for *n* taxa randomly distributed over the taxonomy. While both NRI and NTI increase with increasing clustering they become negative with dispersed taxa. Clustering at terminal nodes causes more increase in NTI relative to NRI. We calculated both measures with respect to the reference taxonomy for each of the 18 groups using 999 randomizations. The corresponding functions were implemented in R (version 2.11.1).

4.2.9 DATA AVAILABILITY

The data used in this study can be obtained from http://algbio.cs.uni-duesseldorf.de/webapps/wa-download/index.php.

4.2.10 DISTANCE METRICS

The distance metrics used for comparison are described below. The metrics were chosen to reflect the diversity of the popular metrics found in the literature, in terms of oligonucleotide lengths, normalization strategies and distance metrics. In the following p denotes the length of the genome signature vectors.

GROUP-SPECIFIC

The group-specific distance between two signatures of genomes from a group is given by;

$$\mathrm{Specific}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2 M_{ii}} \qquad \textbf{Eq. 4.5}$$

Where $\mathbf{M}$ is a diagonal matrix learned by maximizing the estimated generalization performance with training data from the same group (as x and y). For simplicity, the group-specific distance metrics will be referred to as specific distance metrics.

RANDOM LEARNED

The random distance between two signatures calculated for a pair of genomes from a group is given by;

$$\mathrm{Rand}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2 M_{ii}} \qquad \textbf{Eq. 4.6}$$

Where $\mathbf{M}$ is a diagonal matrix learned by maximizing estimated generalization performance using randomly selected training data. For simplicity this metric will be referred to as random metric.

EUCLIDEAN DISTANCE

The Euclidean distance between two signatures is defined as following;

$$\mathrm{Eucl}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p}(\mathrm{x_i} - \mathrm{y_i})^2} \qquad \textbf{Eq. 4.7}$$

This distance was used with the l4n1 and l6n1 signatures.

EUCLIDEAN PCA

This distance was calculated similarly to the Euclidean distance, but in a lower dimensional space after application of principal component analysis (PCA) to retain either the principal components explaining at least one original variable, that is the principal components with

eigenvalue>=1 or three principal components, whichever is larger. This distance metric was used with the l4n1 signature.

DELTA DISTANCES

The delta distance (Mrazek 2009) between two signature vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as following;

$$\mathrm{Delta}(\mathbf{x},\mathbf{y}) = \frac{1}{p}\sum_{i=1}^{p}|x_i - y_i| \qquad \textbf{Eq. 4.8}$$

Here p is number of elements in the vector (256 for tetranucleotide signature). The delta distance between two genomes G1 and G2 was calculated using all pairs of non-overlapping 50 kb segments. If n_1 and n_2 are number of non-overlapping segments X and Y in genomes G1 and G2 respectively then the delta distance between the genomes was calculated as;

$$\mathrm{Delta\,50\,kb}(G1,G2) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathrm{d}_{\mathrm{delta}}(\mathbf{X_i},\mathbf{Y_j}) \qquad \textbf{Eq. 4.9}$$

This distance was used with the l4n1 signature.

CVTREE DISTANCES

The CVTree signature was calculated using oligonucleotides of length 6 normalized by constituent 4- and 5-mers (Gao et al. 2007). The sequences were appended with their reverse complement for calculating this signature. The expected frequency of a hexanucleotide '*abcdef*' was calculated as;

$$\mathrm{fr}^0(abcdef) = \frac{\mathrm{fr}(abcde)\,\mathrm{fr}(bcdef)}{\mathrm{fr}(bcde)} \times \frac{(L-k+1)(L-k+3)}{(L-k+2)^2} \qquad \textbf{Eq. 4.10}$$

Here L is the length of the sequence and k is the length of the oligonucleotides (k=6 for hexanucleotides). Then the normalized elements of the signature vector were then calculated as following;

$$\alpha(abcdef) = \begin{cases} \dfrac{\mathrm{fr}(abcdef) - \mathrm{fr}^0(abcdef)}{\mathrm{fr}^0(abcdef)} & \text{if } \mathrm{fr}^0 \neq 0 \\ 0 & \text{if } \mathrm{fr}^0 = 0 \end{cases} \qquad \textbf{Eq. 4.11}$$

The distances between the resulting vectors were calculated using the cosine similarity.

$$\mathrm{CVTree}(\mathbf{x},\mathbf{y}) = \frac{1-\mathrm{cosine}(\mathbf{x},\mathbf{y})}{2} \qquad \textbf{Eq. 4.12}$$

COMPOSITIONAL SPECTRUM DISTANCES

Compositional spectrum (CompSpec) distances over the DNA alphabet were calculated using the parameter settings as in (Kirzhner et al. 2007b). We first generated 200 random oligonucleotides of length 10 and then counted their imperfect occurrences of up to 2 mismatches (the l10r2 signature) over the complete genomes. The distances between the resulting 200 dimensional vectors were calculated using Spearman's rank correlation coefficient ρ as;

$$\mathrm{CompSpec}(\mathbf{x},\mathbf{y}) = 1 - \rho(\mathbf{x},\mathbf{y}) \qquad \textbf{Eq. 4.13}$$

An important aspect, in our opinion, of the CompSpec is that it only covers a subset of the whole compositional space. For instance, employed parameters account for 9200 ($200*(1+{}^{10}C_2)$) words out of 1048576 (4^{10}) possible words amounting less than 1%. We speculate that the information loss due to this low coverage is, at least partly, responsible for lower performance of CS distances. Although many samples of 200 words are used to build a number of trees which are then aggregated into a final tree using a consensus method (Kirzhner et al. 2007b), it is not straightforward to compare the resulting distances in this way. Therefore we used a single sample of 200 words in this study.

FEATURE FREQUENCY PROFILE DISTANCES

The FFP distances were calculated using the program ffp version 3.19 downloaded from http://ffp-phylogeny.sourceforge.net/. The two-letter RY alphabet was used along with the length of l-mers set to 10. The distance between the normalized feature frequency profile vectors **x** and **y** were calculated using the Jensen-Shannon divergence as;

$$\mathrm{FFP}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathrm{KL}(\mathbf{x},\mathbf{z}) + \frac{1}{2}\mathrm{KL}(\mathbf{y},\mathbf{z}) \qquad \textbf{Eq. 4.14}$$

Here $z_i = (x_i + y_i)/2$ and KL is the Kullback-Liebler divergence.

GENOME BLAST DISTANCES

The whole genome BLAST distances between two genomes were calculated by using the alignments performed by bl2seq program available in the NCBI BLAST executable (version 2.2.18) with default parameters. The resulting tabular report was then parsed using BioRuby (version 1.4.1) (Goto et al. 2010) and the high scoring pairs were converted into a similarity score using the greedy version of the GBDP algorithm without trimming (Henz et al. 2005). Due to computational restrictions we used only one directional alignment instead of averaging over both directions.

$$\mathrm{GBDP}(G1,G2) = -\log\frac{|G1_{match}| + |G2_{match}|}{2\times\min(|G1|,|G2|)} \qquad \textbf{Eq. 4.15}$$

4.2.11 OTHER METHODS

The inferred distance metrics was subsequently used to construct ultrametric trees. Ultrametric trees were inferred with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm of the "phangorn" package in the R statistical environment. Tree topologies were compared to the reference tree topology based on the quartet distance. Principal component analysis (PCA) was performed in R (version 2.11.1) with the "princomp" function. The data was centered and scaled to unit variance before performing PCA.

4.2.12 EXPERIMENTAL SETUP

The tetranucleotide signature corrected for bias in base frequencies (l4n1), i.e. normalized using the zero-order Markov criterion, was chosen to learn the metrics, as it is has been previously shown to contain a strong phylogenetic signal (Pride et al. 2003; van Passel et al. 2006; Mrazek 2009). The Euclidean distance on the l4n1 signature was used as the baseline for comparison. We used two measures to quantify the performance of the methods: The first is the cophenetic correlation coefficient (CPCC) (Sokal & Rohlf 1962) using Spearman's rank correlation, which is also a part of the optimization function used to learn the specific metrics (see Materials and Methods). We also calculated the normalized quartet distance (Nielsen et al. 2011) (referred to as quartet distance hereafter) between two trees built with UPGMA; one using the phenetic distances and the other using the genome-based distances (see section 4.2). We say that a metric performs better only if it shows improvement on both measures; that is a higher CPCC and a lower quartet distance. We show results for 18 groups defined by five different attributes (phylogeny, genomic GC-content, habitat, growth temperature and oxygen requirement, Supplementary Table 4).

For the proposed metric learning method to be of practical value, it is necessary that it is able to learn a generalizable distance metric, a metric that works well on novel genomes not used for learning, from a limited amount of data. Therefore, our experimental setup consisted of randomly sampling genomes of 30 species (one genome per species) from a group and then learning a Mahalanobis metric from the corresponding l4n1 signatures guided by the target phenetic distances such that the estimated generalization performance is maximized (see Materials and Methods). A Mahalanobis metric learned using signatures from one group is referred to as a group-specific metric. The performance of a learned metric was quantified on the test genomes, that is, the genomes from the same group not used for learning the metric. For a set of test genomes, distances were then computed with the learned metric and compared to the corresponding phenetic distances. At the same time the performance of the other methods was also quantified on the test genomes by comparing their distances with the phenetic distances. This procedure was repeated 30 times for each of the 18 groups by using different random training samples, to quantify the variability of the learned metrics. This resulted in 30 performance measurements for the CPCC and quartet distances for each group and each method. Note that for Actinobacteria only 28 metrics were learned due to premature termination of the processes on the computational cluster. The statistical significance of an observed improvement in the 30 repetitions was tested using a one sided Wilcoxon rank sum test. While for CPCC, the alternative hypothesis was that a metric produces higher CPCC values

than the baseline metric, for the QD, the alternative hypothesis was that a metric results in lower quartet distances than the baseline metric.

4.3 RESULTS

4.3.1 PHYLUM

We begin by showing that the taxonomic signal of the l4n1 genomic signature can be improved with specifically learned metrics for phylogenetic groups at the phylum level. Four extensively sequenced phyla, the Proteobacteria, Firmicutes, Actinobacteria and Euryarchaeota, were chosen for this analysis (Supplementary Table 4).

Our results show that better distance metrics, that is higher cophenetic correlation and lower quartet distance on the test genomes when compared to the baseline, could be learned for the phylogenetic groups except for Euryarchaeota, where the learned metrics did not show improvement over the Euclidean metric (Figure 4.1, Supplementary Table 6). The Proteobacteria metrics showed only marginal but significant ($P<0.05$, Wilcoxon test) improvement, which might be due of its diverse and non-monophyletic nature (Garrity 2005). Such disagreement with taxonomy was also observed with Proteobacterial CVTree based on translated protein products (Li, Xu & Hao 2010). The best performance improvement due to specific metrics was observed for the phylum Actinobacteria, where the average cophenetic correlation significantly increased from 0.39 to 0.64 ($P=8.23e-10$, Wilcoxon test) while the average quartet distance decreased from 0.53 to 0.43 ($P=2.73e-13$, Wilcoxon test). More than 25 (out of the 30) learned metrics showed significantly different correlation coefficients for the Proteobacteria, Firmicutes and Actinobacteria (Hotelling-Williams test, $P<0.05$) (Supplementary Figure 11). The other l4n1 based distances, the Euclidean distances after applying PCA and the delta distances averaged over 50 kb segments, performed either similar or only slightly better than the baseline. The metrics learned from randomly sampled species over the entire taxonomy performed worse than the baseline except for a slight performance improvement for the Actinobacteria. The phyla-specific metrics also performed better than the l6n1 signature-based Euclidean distances. This shows the advantage of learning specific metrics in comparison to signatures based on longer oligonucleotides.

The phyla-specific metrics also performed better than the l6n1 signature-based Euclidean distances. This shows the advantage of learning specific metrics in comparison to signatures based on longer oligonucleotides. The Euclidean distances based on the l4n1 and l6n1 signatures performed similarly, except for the Actinobacteria, where the l6n1 signature performed better. CVTree with the l6n5,4 signature showed overall better performance than the l6n1 Euclidean distances, the compositional spectrum and FFP distances performed less well in comparison. Interestingly, all signature-based distances with long oligonucleotides (Eucl l6n1, CVTree l6n5,4, CompSpec l10r2 and FFP l10ry) with lower overall cophenetic correlation, except for FFP, performed better for the Actinobacteria than the baseline ($P<0.05$, Wilcoxon test). This might be due to the close relatedness of the genomes in the phylum Actinobacteria and their characteristically high GC-content, making longer oligonucleotides more informative. For all groups except Firmicutes, the alignment-based method GBDP performed better than alignment-free methods, however, this comes at a considerable computational cost.

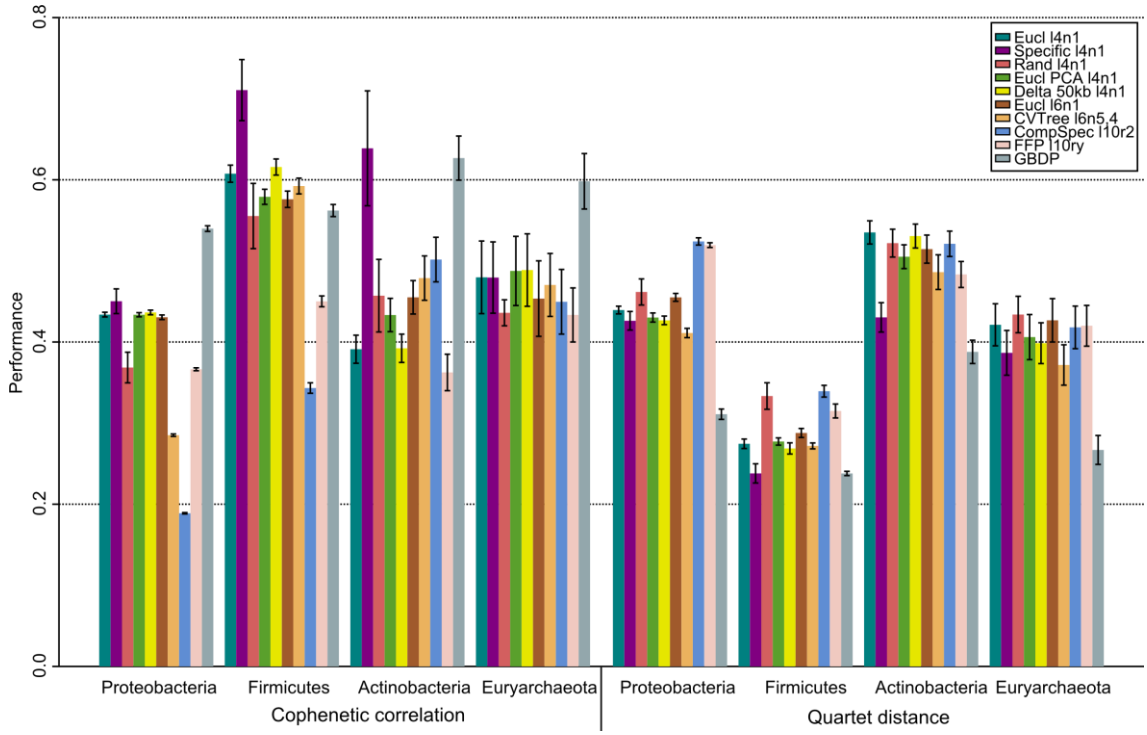


Figure 4.1. Performance on the phylogenetic groups. Each bar shows performance measure along with error bars showing standard deviation.

4.3.2 GC-CONTENT

We performed similar experiments with the genomes divided into three groups according to their genome-wide GC-content (<=30%, >30%-<=50% and >50%-<=70%, Supplementary Table 4). It has been previously noted that GC-content affects oligonucleotide based trees grouping similar GC-content genomes together irrespective of their phylogenetic relationships and tetra to octanucleotide frequency based trees of genomes with similar GC-content show high congruence with gene based trees at genus and family level (Takahashi et al. 2009). Therefore we expected that improved distance metrics could be learned for groups of genomes with similar GC-content. The GC-specific metrics we inferred improved in cophenetic correlation over the baseline for all three GC-content groups.

There was also a decrease in the quartet distance for the genomes with 30% or less GC-content and for genomes with GC-content between 50% and 70% (Figure 4.2). Most metrics for the individual groups had significantly different correlation coefficients from the baseline method ($P<0.05$, computed with Hotelling-Williams test) (Supplementary Figure 11). In general while a strong signal was observed for all the alignment-free methods on the low GC-content group, a weaker signal was observed on the moderate GC-content genomes (Figure 4.2, Supplementary Table 6). Of the other alignment-free methods, only CVTree consistently and significantly ($P<8.2e-6$, Wilcoxon test) performed better than the baseline. The compositional spectrum and FFP methods performed well only on the genomes with GC-content of 30% or less. GBDP performed better than the baseline in all the groups and performed worse than the learned l4n1 metrics on the 30% or less GC-content group.

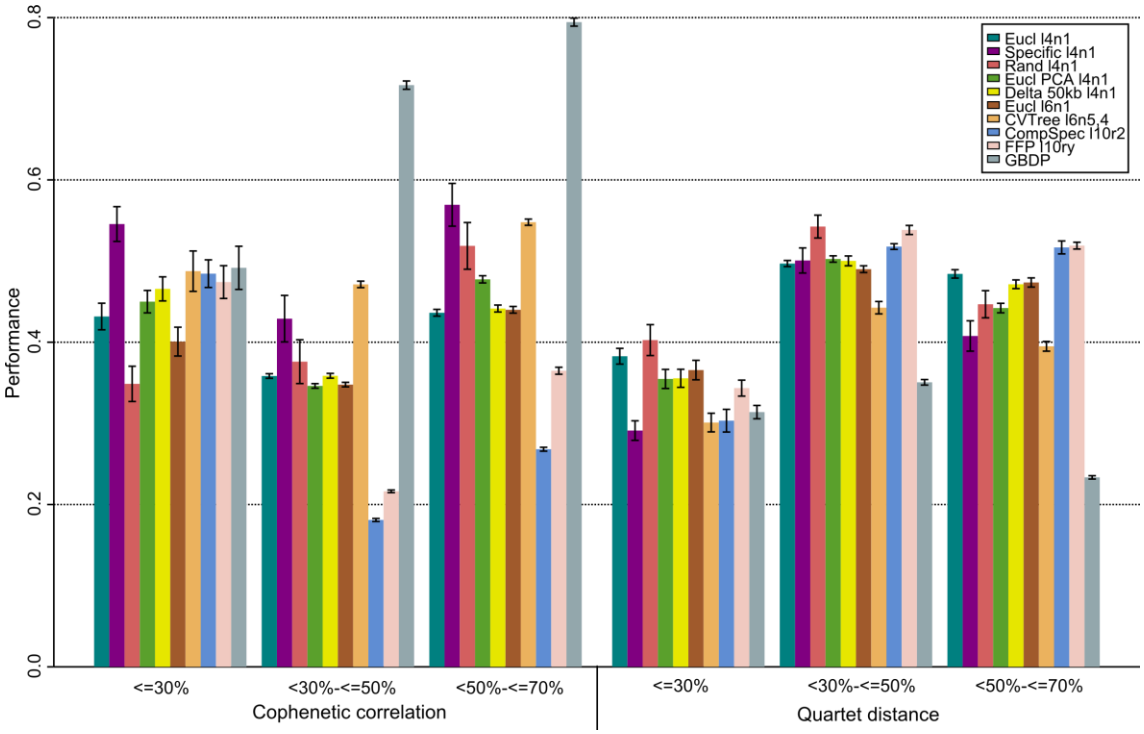


Figure 4.2. Performance on the GC-content groups. Each bar shows performance measure along with error bars showing standard deviation.

4.3.3 ECOLOGICAL ATTRIBUTES

Next we investigated whether specific metrics for ecological groups show an improvement over the baseline. This is a challenging task as ecological groups might contain distantly related genomes, a scenario in which alignment-free methods can face difficulties (Mrazek 2009). Three ecological factors were chosen to define groups: habitat (5 groups), temperature range (3 groups) and oxygen requirement (3 groups) (Supplementary Table 4).

The habitat-specific l4n1 metrics showed an improvement over the baseline both in terms of the CPCC and the quartet distance for all five groups. Only the improvement of the quartet distance for the host-associated metrics was not significant (Figure 4.3, Supplementary Table 6). While CVTree showed an increase in the CPCC for all five habitat groups, but also an increased quartet distance for the aquatic and specialized groups, FFP showed an improvement over the baseline only for the multiple habitat genomes (P<7.74e-15, Wilcoxon test).

In computation of taxonomic distances and genome trees for genomes from all three temperature range groups, the learned l4n1 metrics performed better than the baseline (P<7e-3, Wilcoxon test), except for an increase in the quartet distance for the mesophiles group. Interestingly, for the mesophiles group 19 specific metrics did show a significant change in correlation (Supplementary Figure 11). CVTree performed well for all groups except for a decrease in the CPCC for hyperthermophiles, while FFP showed improvement only for the hyperthermophiles group (P<1.3e-3, Wilcoxon test).

We also observed an improvement for the learned l4n1 metrics for all oxygen-requirement types (aerobe, anaerobe and facultative anaerobes) (P<1.2e-6, Wilcoxon test), except for a performance reduction in term of an increase in the quartet distance for the facultative anaerobes. CVTree, as before, showed improvement for the anerobes and facultative groups (P<3.15e-15, Wilcoxon test) and performed similarly to GBDP for the genomes of the facultative anaerobes. While Euclidean metric on the l4n1 signature after performing PCA showed marginal but significant improvement for aerobes and anaerobes, Delta50kb and Euclidean metric on the l6n1 signature showed significant improvements for anaerobes and facultative groups, respectively. The other methods did not show a consistent performance pattern.

Overall, for all eleven ecological groups 23 or more metrics showed significant change in correlation coefficients with the phenetic metric of the reference taxonomy in comparison to the baseline (P<0.05, computed with Hotelling-Williams test). For three habitats - aquatic, host-associated and specialized – as well as the mesophilic and aerobic groups, all 30 metrics differed significantly (Supplementary Figure 11). GBDP performed best for all groups defined by the three ecological attributes (P<1.46e-9, Wilcoxon test).

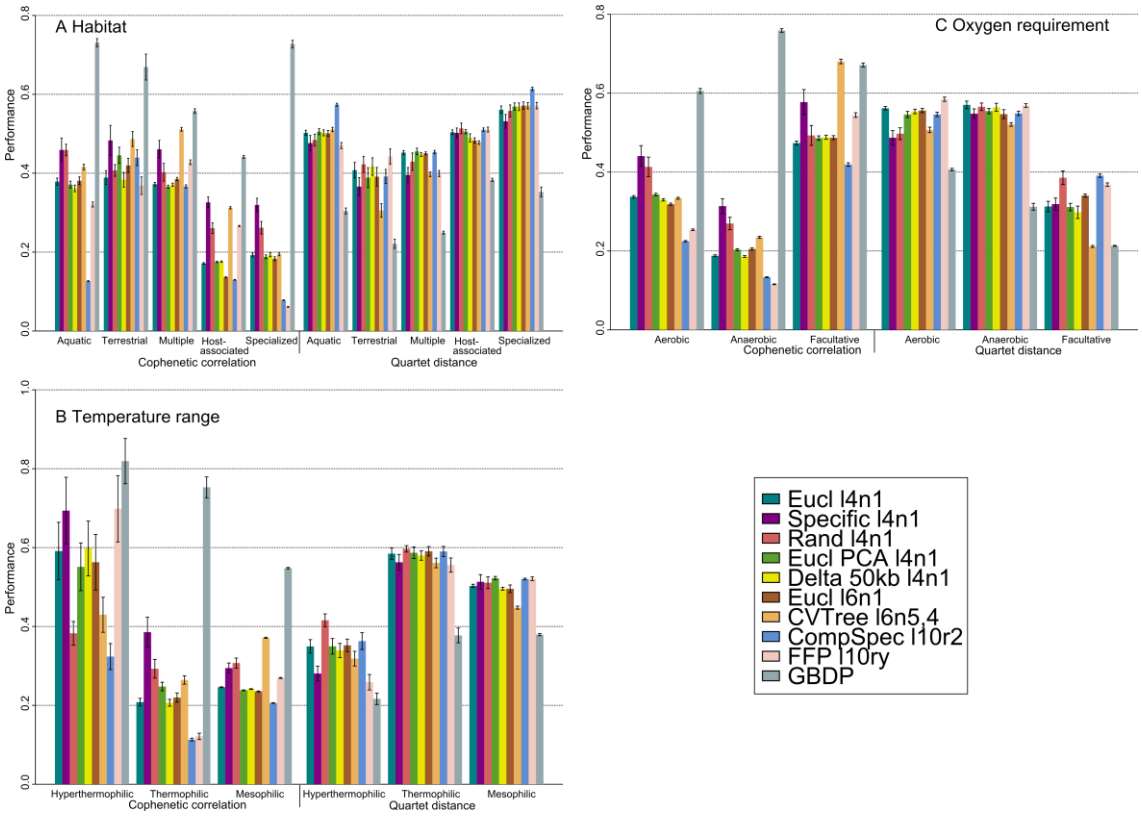


Figure 4.3. Performance on the ecological groups from three attributes. The bars show performance measures and the error bars indicate standard deviation.

4.3.4 GROUP-SPECIFIC METRICS NOTABLY IMPROVED TREE INFERENCE

One could argue that a learned metric performs well for a group by chance and not because it inferred specifics of evolutionary rates for different tetranucleotides for the group. To investigate this question we learned 30 metrics from 30 randomly selected species each

(referred to as random metrics hereafter) and compared their performance to the performance of the 30 group-specific learned metrics for each of the 18 groups with one sided Wilcoxon signed rank sum test. We tested whether the group-specific metrics produce higher CPCC and lower quartet distance than the random metrics. Note that the random metrics showed significantly better performance with respect to the baseline metric for Actinobacteria, GC content between 50% and 70%, aquatic and aerobic groups (P<3.61e-2, Wilcoxon test) (Supplementary Table 6)

Table 4.1. P-values from one-sided Wilcoxon signed rank sum tests to check specificity of the learned metrics to their respective groups. While for CPCC the alternative hypothesis was that the group-specific metrics produce higher CPCC values than randomly learned metrics, for QD the alternative hypothesis was that the group specific metrics produce lower quartet distance than randomly learned metrics. Significant values (<0.05) are shown in boldface.

Attribute	Group	CPCC	QD
Phylum	Proteobacteria	**0.0000**	**0.0001**
	Firmicutes	**0.0000**	**0.0000**
	Actinobacteria	**0.0000**	**0.0000**
	Euryarchaeota	**0.0032**	**0.0029**
GC-content	<=30%	**0.0000**	**0.0000**
	>30%-<=50%	**0.0014**	**0.0000**
	>50%-<=70%	**0.0000**	**0.0013**
Habitat	Aquatic	0.5957	0.3762
	Terrestrial	**0.0000**	**0.0005**
	Multiple	**0.0000**	**0.0057**
	Host-associated	**0.0000**	**0.0386**
	Specialized	**0.0001**	**0.0006**
Temperature range	Hyperthermophilic	**0.0000**	**0.0000**
	Thermophilic	**0.0001**	**0.0000**
	Mesophilic	0.8850	0.6349
Oxygen requirement	Aerobic	**0.0154**	0.1150
	Anaerobic	**0.0030**	**0.0011**
	Facultative	**0.0000**	**0.0000**

For all the groups, except aquatic, mesophiles and aerobes, the specifically learned metrics performed significantly better than the random metrics (P<3.86e-2, Wilcoxon test) (Table 4.1). This implies that the group-specific metrics perform better than the ones learned on randomly sampled genomes and group-specific aspects of tetranucleotide usage allow an improved inference of the taxonomic relationships for the respective organisms. The lack of improvement for aquatic species, mesophiles and aerobes might be in part caused by abundance of these groups among the genomes (Supplementary Table 6). This may have resulted in some of the learned metrics from randomly selected species to partially represent specific properties of these groups.

4.3.5 DIMENSIONALITY REDUCTION RESULTED IN MARGINAL IMPROVEMENT

Unsupervised dimensionality reduction techniques, such as principal component analysis (PCA), have been used for noise reduction and visualization of genome signatures (Sandberg et al. 2001; Mrazek 2009). PCA embeds the input space into a potentially lower dimensional space defined by orthogonal basis vectors. Inferring taxonomic distances based on Euclidean distances after applying PCA to l4n1 signatures mostly resulted in a marginal or no improvement (Figure 4.1, Figure 4.2 and Figure 4.3). The marginal improvement result is interesting, as it suggests the existence of lower dimensional genomic signature space.

Table 4.2. Cophenetic correlation coefficient and quartet distance before (CPCC, QD) and after (CPCC_PCA, QD_PCA) principal component analysis. The dimension and variance columns show number of dimensions and variance retained respectively. No significant improvement was observed after applying PCA either for the CPCC or the QD (P>0.3, one-sided Wilcoxon rank sum test).

Attribute	Group	CPCC	CPCC_PCA	QD	QD_PCA	Dimension	Variance (%)
Phylum	Proteobacteria	0.42	0.43	0.45	0.43	21	94
	Firmicutes	0.57	0.54	0.32	0.29	20	96
	Actinobacteria	0.39	0.44	0.55	0.50	19	96
	Euryarchaeota	0.46	0.45	0.47	0.43	20	97
GC-content	<=30%	0.30	0.34	0.43	0.40	19	97
	>30%-<=50%	0.36	0.34	0.51	0.51	25	94
	>50%-<=70%	0.44	0.48	0.48	0.43	22	94
Habitat	Aquatic	0.39	0.38	0.51	0.51	24	95
	Terrestrial	0.39	0.45	0.39	0.38	18	96
	Multiple	0.37	0.36	0.46	0.45	21	95
	Host-associated	0.17	0.18	0.51	0.51	21	95
	Specialized	0.20	0.19	0.57	0.57	23	95
Temperature range	Hyperthermophilic	0.46	0.41	0.43	0.46	18	98
	Thermophilic	0.19	0.24	0.59	0.58	22	96
	Mesophilic	0.25	0.24	0.51	0.52	22	93
Oxygen requirement	Aerobic	0.34	0.34	0.56	0.56	22	95
	Anaerobic	0.19	0.20	0.58	0.55	24	95
	Facultative	0.46	0.47	0.30	0.35	23	95

To further investigate this effect, we calculated cophenetic correlations and quartet distances for all the groups individually to l4n1 distances with and without using PCA (Table 4.2). The dimensionality of the reduced space was selected to be the dimensions explaining at least one original variable, i.e. dimensions with eigenvalues of at least one. Interestingly, approximately 20 dimensions (18-25) were retained for all the groups capturing 93-98% of variance. Although PCA resulted in a marginal non-significant (P>0.3, Wilcoxon test) improvement it performed less well than the group-specific metrics. Similarly, when PCA was applied to the l6n1 signature with the Euclidean distance metric, a high reduction in the dimensionality was observed (38-

114 principal components explaining 97.81-99.96% variance), with no significant (P>0.25, Wilcoxon test) performance improvement (Supplementary Table 7).

4.3.6 TRENDS ACROSS GROUPS

We investigated whether the genomic and taxonomic composition of the groups are relevant for the improvement obtained by the specific metrics over the baseline. The aim of this analysis was to get a better understanding of when application of the proposed method might be most relevant. We calculated nine statistics for the groups (number of genomes, number of species, mean genome size, standard deviation of genome sizes, mean GC-content, standard deviation of GC-content, NRI and NTI) and correlated them with the change in the mean cophenetic correlation of the specific metrics relative to the baseline (Table 4.3, Supplementary Table 5) across the groups. The positive correlation here means that an increase in the statistic corresponds to an improvement in the CPCC on average and vice versa. The Actinobacteria and Euryarchaeota groups were removed from this analysis because they behaved like an outlier with respect to change in the CPCC, above the 99th quartile and below the 1st quartile, respectively.

Table 4.3. Correlation of the mean change in the cophenetic correlation coefficient with different statistics across the groups. Here mean and sdev are average and standard deviation values, NRI and NTI stand for net relatedness index and nearest taxon index respectively. The Actinobacteria and Euryarchaeota groups were removed for this analysis as they behaved like outliers. Significant values (P<0.05) are shown in boldface.

Correlation	Value	#genomes	#species	Genome size (mean)	Genome size (sdev)
Pearson	R	**-0.54**	-0.17	-0.34	-0.33
	P-value	**0.03**	0.52	0.19	0.22
Spearman	ρ	-0.46	-0.13	-0.44	-0.44
	P-value	0.07	0.63	0.09	0.09
Correlation	**Value**	**GC-content (mean)**	**GC-content (sdev)**	**NRI**	**NTI**
Pearson	R	0.03	0.02	**-0.54**	-0.35
	P-value	0.92	0.95	**0.03**	0.19
Spearman	ρ	0.06	0.03	-0.4	-0.26
	P-value	0.81	0.93	0.12	0.32

The strongest and significant negative correlation, Pearson's R=-0.54, P=0.03, was with the phylogenetic community measure net relatedness index (NRI) (Webb et al. 2002). NRI measures the phylogenetic clustering behavior of the taxa; therefore, this negative correlation suggests that as the taxa become more clustered on the taxonomy, the specific metrics provide less improvement. This result was expected, as for closely related taxa the baseline (l4n1 signature with Euclidean distance) is expected to perform well (Mrazek 2009). A lower

and non-significant, but also negative correlation was observed for the nearest taxa index (NTI) (Webb et al. 2002), which increases more if taxa cluster at the terminal nodes. The overall number of genomes in a group also showed a significant negative correlation with the mean change of the cophenetic correlation, suggesting that our method provides a larger improvement in the CPCC for larger groups and groups with bigger genomes. As larger groups are normally more diverse, the baseline performs poorly and an improvement can be achieved with the specific metrics. For the negative correlation with genome sizes we speculate that larger genomes may exhibit a noisy genomic signature, for example due to presence of phages and plasmids (Suzuki et al. 2010), the specific metrics might provide an improvement by learning appropriate weights for oligonucleotides, such that the noise is reduced.

Interestingly, no significant correlation was observed with either the mean or the standard deviation of the GC-content for each group, suggesting that the improvement provided by the specific metric does not depend on the group GC-content, except for the Actinobacteria. Taken together, this analysis suggests that our method provides relatively more improvement when the baseline is expected to perform worse and less improvement otherwise.

4.3.7 The learned group-specific metrics generalized across larger taxonomic distances

To investigate the effect of the genome relatedness on learning group-specific metrics we removed genomes of the same species and order as the ones used for learning independently for each group-specific metric and recomputed the performance measures. These experiments were performed on the 1951 genomes obtained from NCBI GenBank in June 2012. We observed similar trends as before (Supplementary Figure 12-16), suggesting that metric learning is advantageous even when closely related genomes are not available for training. However, in many cases performance of all the tested methods degraded after this removal, indicating that signature based methods indeed perform better at lower taxonomic distances.

4.4 Conclusions

In this work we proposed a method to learn taxonomic distance metrics from genome signatures and the corresponding phenetic distances between them. Our aim was to improve genome signature-based genome tree inference for groups of genomes where the groups were defined by phylogenetic, genomic or ecological attributes. Our empirical analyses showed that genome trees inferred from genome signatures can be improved by learning group-specific distance metrics. As expected, metrics learned for different phyla and GC-content groups showed significant improvement in the quality of inferred genome trees (for three groups out of four and two groups out of three, respectively). Working with the hypothesis that environmental selective forces can shape the nucleotide composition of genomes, that is different niches drive the oligonucleotide composition in different directions, we learned specific metrics for different ecological groups. These ecological group-specific metric showed performance improvement for eight out of eleven ecological groups.

The performance improvement shown by specific metrics for phylogenetic and GC-content groups of species was relatively higher and generalized better for distant genomes than for the

ecological groups. Nevertheless, also for the ecological groups, the learned metrics in most cases showed a performance improvement. The ecological groups in particular contain genomes of species only distantly related to each other, where the alignment-free methods are known to be less accurate. Of the other alignment-free methods evaluated here only CVTree showed a consistent improvement over the baseline. The better performance of CVTree compared to the l6n1 signature might be due to a more appropriate normalization.

For the FFP metric we also computed distances between randomly sampled 50 kb continuous segments from the genomes in order to check whether different sizes of genomes might be confounding the distance calculations. The results were similar (data not shown). We did not implement the block-FFP and optimal range finding algorithms (Sims et al. 2009) and it will be interesting to see whether those lead to performance improvement, but it is out of the scope of this work. Furthermore, our experiments show that dimensionality reduction with PCA does not provide a consistent performance improvement.

An important observation from our analysis was that the BLAST alignment-based genome dissimilarity metric (GBDP) was the overall best performing method, both in terms of the cophenetic correlation and the quartet distance. The good performance of GBDP implies that the information necessary for tree inference can be uncovered using genome-wide alignments. The comparatively lower performance of the alignment-free methods suggest that the distances calculated from the genome signatures do not represent universal taxonomic relationships with the same accuracy. The good performance of GBDP might also partly be due to the use of an evolutionary model. At the same time, the lower performance of alignment-free methods might result from the loss of information while encoding a longer sequence by means of shorter oligonucleotides. Further research is needed to pin point the advantages and shortcomings of the different methods.

However, performing alignments is computationally expensive and hence difficult to scale to a large number of genomes. The group-specific metrics we introduced can be learned from a small number of genomes, i.e. 30 different species, and knowledge of the target phenetic distances in the reference taxonomy. Therefore, to save computational cost, in case a resolved taxonomy for a group of genomes is not available, one could first infer a partial taxonomy from a subset of the genomes with an accurate method like GDBP and then use this partial taxonomy to learn a signature-based group-specific distance metric that in turn could be applied to infer taxonomic distances between the remaining genomes.

In summary, our findings suggest that different types of organisms have specific distance metrics over the genome signature and that these can be uncovered by considering their ecological, genomic or phylogenetic attributes. Our new method performed significantly better than a baseline technique for 13 out of 18 groups, indicating that group-specific aspects define the genome signature and that their consideration can improve the inference of taxonomic relationships. The existence of ecology specific metrics strengthens the hypothesis that environmental factors affect the oligonucleotide usage of genomes. We also repeat the need for more fine grained terms to describe specific environments and sample source information in public repositories, as provided by the environmental ontology (Hirschman et al. 2008). With the rapid advance in sequencing technologies large number of genome from

microorganisms, even the ones not cultivable with traditional sequencing methods, will become available in the near future. Accurate and efficient methods are necessary to analyze this large scale data. Our proposed method is a step towards this goal.

The analysis of the group-specific oligonucleotide weights and whether they provide insights into any characteristics of the group will be an interesting direction for future work. In this work the group-specific metrics were learned only from group-specific data, therefore the learned oligonucleotide weights do not necessarily contain discriminatory information. Furthermore, the limited number of genomes (30) used for learning a metric, in combination with correlations between the oligonucleotides can lead to divergent metrics for a group, where weights can be distributed across different correlated oligonucleotides to obtain the same result, which makes the interpretation of a biological or evolutionary meaning of the learned weights complicated.

5 CONCLUSIONS AND OUTLOOK

In the following sections we present a brief summary of the main conclusions of the work carried out in this thesis. The work done in this thesis is a step forward towards solving the addressed problems, though challenges remain, therefore we also discuss some possible directions for future research.

5.1 CONCLUSIONS

Genomics will play an increasingly larger role in medicine, energy and many other important biotechnological applications. The advent of sequencing technologies means more sequence data being generated than can be efficiently processed using the currently available computational resources. Therefore, devising efficient algorithms that can tackle the large amount of genomic data in a reasonable time is important, should the pace of the genomic sciences as a whole and the benefits it provides be maintained.

To this end, this thesis proposes novel methods to address two important bioinformatics problems; taxonomic assignment of metagenome sequences and inference of genome trees. Both methods rely on the genome signature paradigm for sequence comparison. Genome signatures have two main advantages when used for sequence comparison. Firstly, they allow computationally efficient comparison between genomic sequences, as alignment is not necessary. Secondly, due to their pervasiveness, only segments of genomes are sufficient. Furthermore, both methods are based upon state-of-the-art machine learning methods.

By exploiting the properties of the genome signature along with the use of structural support vector machines we proposed a new method, PhyloPythiaS, for taxonomic assignment of metagenome sequences, an important step in metagenome analyses. Empirical analysis of several simulated and real metagenome sequence samples showed that PhyloPythiaS performs well, especially when only few data from dominant populations are available. Evaluation on simulated and real data showed that PhyloPythiaS performs quite well and outperforms other methods in realistic scenarios. We also evaluated PhyloPythiaS on the contigs or scaffolds from three sequencing technologies resulting in consistently good performance. Furthermore, at assignment time, PhyloPythiaS is considerably faster than other methods, which will facilitate analysis of large metagenome samples.

The structural SVM used for taxonomic assignment needs a reference hierarchy describing relationships between the taxa. Currently we use the reference taxonomy from NCBI. In future, with a large number of genome sequences produced, direct generation of a hierarchy from the genomes will be useful and therefore we explored the use of genome signature to infer genome trees. We developed a metric learning method to infer taxonomic distances between genomes based on the genome signature. A primary hypothesis was that different taxonomic distances between groups of genomes, defined by phylogenetic, GC-content and ecological factors, are better defined by group-specific metrics. Empirical analysis of 18 groups showed that the proposed method performs well on most of the groups.

5.2 OUTLOOK

This work was confined to the use of linear kernels and it will be interesting to explore performance when non-linear or sequence alignment kernels (Watkins 2000) are used. The decision to use of linear kernel was primarily due to higher computational cost incurred by use of kernels, particularly for the structural SVM, where the training time computational complexity is linear in the number of examples for linear kernel, it scales quadratically for other kernels (Joachims et al. 2009). Therefore, use of kernels is impractical for large data sets as used in this thesis. Two possible directions can be followed as a remedy; sparse approximation of the kernel matrices (Joachims et al. 2009) and use of faster optimization techniques. Use of alternative formulations of the structural SVM (Sarawagi & Gupta 2008) can also lead to more accurate results. Furthermore, as available sequence data and the breadth of taxonomy grow, the training phase of the structural SVM can become an issue. Towards this end, incremental techniques that reuse existing solutions while learning new models incorporating more sequence data and a larger hierarchy in order to reduce execution time will be extremely useful. Another issue to tackle in the future is the lower performance of alignment-free methods for assignment of short (<1000 bp) sequences. This is due to the limitations on the pervasiveness of the genome signature and therefore difficult to solve. Currently, sequence assemblies are used to obtain longer sequences in order to circumvent this issue. As sequencing technologies progress, the increased read length will automatically offer a solution.

In the case of genome tree inference problems the current work was confined to learning linear distance metrics. This can be extended to learning non-linear distance metrics in the future, which may lead to further performance improvements. It will be also interesting to check whether learning a full matrix instead of a diagonal matrix proves to be beneficial. We here used the cophenetic correlation with Spearman's rank correlation coefficient as the objective function. Although, the increase in the cophenetic correlation was correlated with the decrease in the quartet distance (Pearson's R=0.46, P<2.2e-16; all the 18 groups combined), further research might identify other suitable optimality criteria. Furthermore, distance metric learning might be extended to unsupervised binning of metagenome data (McHardy & Rigoutsos 2007) in order to improve performance on a particular ecological niche, such as the human-gut.

6 SUPPLEMENT

6.1 SUPPLEMENTARY TABLES

Supplementary Table 1. Modeled taxa for the TW sample. Only the leaf taxa are shown, all the clades at more general taxonomic ranks were included in the modeled taxonomy.

NCBI scientific name	NCBI taxonomic identifier	Sample-specific data (kb)
Acinetobacter	469	--
Actinobacteria (class)	1760	--
Bradyrhizobiaceae	41294	--
Campylobacter	194	--
Desulfovibrionaceae	194924	--
Enterobacteriaceae	543	--
Eubacteriaceae	186806	--
Fusobacteriaceae	203492	--
Methanomicrobiales	2191	--
Methanosarcina	2207	--
Pasteurellaceae	712	--
Prevotellaceae	171552	--
Psychrobacter	497	--
Ruminococcaceae	541000	--
Selenomonas	970	--
Staphylococcus	1279	--
Thermoplasma	2302	--
uncultured Erysipelotrichaceae bacterium (WG-3)	331630	5.7
uncultured Lachnospiraceae bacterium (WG-2)	297314	143
uncultured Succinivibrionaceae bacterium (WG-1)	538960	257

Supplementary Table 2. Number of contigs classified by different methods at different taxonomic ranks for the TW sample. Out of the 5,995 contigs in total for this metagenome sample. All numbers indicate the raw output of every method. PhyloPythia does not classify fragments shorter than 1,000 bp so the total number of contigs classified is less (5,245).

Taxonomic rank	PhyloPythiaS	PhyloPythia	PhymmBL	MEGAN
Domain	1,206	1,579	--	630
Phylum	503	485	--	191
Class	214	261	92	85
Order	1,748	801	1,086	401
Family	997	1,012	250	288
Genus	71	--	2,899	1,446
Species	1,255	1,062	1,525	277
Not assigned	1	45	143	2,677

Supplementary Table 3. Modeled clades for PhyloPythiaS for the human gut metagenome samples (TS28 and TS29). Only the leaf clades are shown, all the clades at more general taxonomic ranks were included in the modeled taxonomy. Only part of the sample-specific data was used to learn PhyloPythia and PhyloPythiaS models (see Supplementary notes).

NCBI scientific name	NCBI taxonomic identifier	Sample-specific data (kb)
Alistipes	239,759	198
Anaerococcus	165,779	1,300
Anaerotruncus	244,127	74
Atopobium	1,380	--
Bacteroides	816	23,600
Bifidobacterium	1,678	3,800
Blautia	572,511	13
Butyrivibrio	830	6.2
Clostridium	1,485	7,200
Collinsella	102,106	512
Coprococcus	33,042	29
Dorea	189,330	1,500
Escherichia	561	--
Eubacterium	1,730	600
Faecalibacterium	216,851	2,300
Finegoldia	150,022	--
Holdemania	61,170	7.7
Lactococcus	1,357	--
Methanobrevibacter	2,172	1,300
Methanothermobacter	145,260	--
Parabacteroides	375,288	1,600
Porphyromonas	836	--
Providencia	586	--
Roseburia	841	31
Ruminococcus	1,263	4,000
Streptococcus	1,301	--

Supplementary Table 4. Taxonomic breakdown of the 18 groups comprising five attributes.

Attribute	Group	Genomes	Species	Genus	Family	Order	Class	Phylum	Domain
Phylum	Proteobacteria	507	335	180	71	36	6	1	1
	Firmicutes	199	109	43	23	6	2	1	1
	Actinobacteria	91	76	45	33	6	1	1	1
	Euryarchaeota	53	49	34	16	11	9	1	1
GC-content	<=30%	77	51	22	14	12	9	7	2
	>30%-<=50%	505	337	171	94	65	34	25	2
	>50%-<=70%	458	332	207	107	64	28	17	2
Temperature range	Hyperthermophilic	47	41	22	13	11	9	7	2
	Thermophilic	70	68	56	40	30	22	18	2
	Mesophilic	830	546	293	142	75	36	23	2
Habitat	Aquatic	169	143	113	69	51	27	18	2
	Terrestrial	71	63	46	40	23	16	9	2
	Multiple	294	188	112	72	44	18	10	2
	Host-associated	330	209	101	62	37	20	13	2
	Specialized	115	106	86	52	46	31	19	2
Oxygen requirement	Aerobic	331	255	165	98	52	24	18	2
	Anaerobic	198	169	112	69	45	31	19	2
	Facultative	345	199	92	50	35	17	12	2

Supplementary Table 5. Group statistics. The table is divided in two parts for convenience. Here mean and sdev are average and standard deviation values. NRI and NTI stand for net relatedness index and nearest taxon index respectively.

A.

Attribute	Group	Change_rho	#organisms	#species	Genome size (mean)	Genome size (sdev)
Phylum	Proteobacteria	0.02	507	335	4066140	1853855
	Firmicutes	0.10	199	109	3098486	1241844
	Actinobacteria	0.25	91	76	4613913	2262225
	Euryarchaeota	0.00	53	49	2378586	918270.2
GC-content	<=30%	0.11	77	51	1703760	1360259
	>30%-<=50%	0.07	505	337	2846184	1450872
	>50%-<=70%	0.13	458	332	4417507	1708588
Habitat	Aquatic	0.08	169	41	3446547	1516513
	Terrestrial	0.09	71	68	5500797	2205847
	Multiple	0.09	294	546	4220323	1674874
	Host-associated	0.15	330	143	2809106	1781657
	Specialized	0.13	115	63	2676180	1279979
Temperature range	Hyperthermophilic	0.10	47	188	2028211	510843.1
	Thermophilic	0.18	70	209	2705110	1172844
	Mesophilic	0.05	830	106	3722610	1913293
Oxygen requirement	Aerobic	0.10	331	255	4217805	2192102
	Anaerobic	0.13	198	169	2855070	1237975
	Facultative	0.10	345	199	3677349	1601115

B.

Attribute	Group	GC-content (mean)	GC-content (sdev)	Z-score	NRI	NTI
Phylum	Proteobacteria	52	12.1	1.53	35.14	9.65
	Firmicutes	38	6.9	2.07	27.59	10.66
	Actinobacteria	65	6.9	1.82	23.77	5.37
	Euryarchaeota	47	12.2	1.23	7.75	4.52
GC-content	<=30%	27	2.4	3.03	3.29	8.16
	>30%-<=50%	40	5.3	0.82	1.01	6.87
	>50%-<=70%	60	6.2	1.03	8.92	3.06
Habitat	Aquatic	44	8.8	0.27	2.27	1.14
	Terrestrial	49	12.7	2.39	1.03	1.60
	Multiple	49	13.3	1.13	5.11	7.27
	Host-associated	49	11.3	1.35	2.87	8.46
	Specialized	59	13.2	1.04	-3.85	0.43
Temperature range	Hyperthermophilic	50	12.8	3.70	0.33	4.93
	Thermophilic	44	12.8	1.19	-1.06	0.85
	Mesophilic	48	12.7	-0.13	6.53	4.65
Oxygen requirement	Aerobic	54	14.1	0.58	1.21	1.17
	Anaerobic	45	11.6	0.88	-5.06	1.57
	Facultative	47	10.9	1.69	11.41	10.62

Supplementary Table 6. P-values of one-sided Wilcoxon signed rank sum tests to check improvement of different methods over the baseline Euclidean l4n1 method. While for CPCC the alternative hypothesis was that a metric produces higher CPCC values than the baseline metric (A), for QD the alternative hypothesis was that a metric produces lower quartet distance than the baseline metric (B). Significant values (<0.05) are shown in boldface.

A.

Attribute	Group	Specific l4n1	Rand l4n1	Eucl PCA l4n1	Delta 50kb l4n1	Eucl l6n1	CVTree l6n5,4	CompSpec l10r2	FFP l10ry	GBDP
Phylum	Proteobacteria	**0.017**	1.000	0.509	0.053	0.974	1.000	1.000	1.000	**0.000**
	Firmicutes	**0.000**	1.000	1.000	**0.014**	1.000	0.999	1.000	1.000	1.000
	Actinobacteria	**0.000**	**0.000**	**0.000**	0.497	**0.000**	**0.000**	**0.000**	0.987	**0.000**
	Euryarchaeota	0.462	0.988	0.322	0.291	0.864	0.719	0.945	0.989	**0.000**
GC-content	<=30%	**0.000**	1.000	**0.017**	**0.000**	0.998	**0.000**	**0.000**	**0.000**	**0.000**
	>30%-<=50%	**0.000**	0.067	1.000	0.456	1.000	**0.000**	1.000	1.000	**0.000**
	>50%-<=70%	**0.000**	**0.000**	**0.000**	**0.021**	**0.045**	**0.000**	1.000	1.000	**0.000**
Habitat	Aquatic	**0.000**	**0.000**	0.885	0.997	0.327	**0.000**	1.000	1.000	**0.000**
	Terrestrial	**0.000**	0.067	**0.000**	0.651	**0.006**	**0.000**	**0.000**	0.888	**0.000**
	Multiple	**0.000**	**0.004**	0.993	0.719	**0.000**	**0.000**	0.985	**0.000**	**0.000**
	Host-associated	**0.000**	**0.000**	0.096	0.060	1.000	**0.000**	1.000	**0.000**	**0.000**
	Specialized	**0.000**	**0.000**	0.801	0.485	0.927	0.421	1.000	1.000	**0.000**
Temperature range	Hyperthermophilic	**0.002**	1.000	0.914	0.462	0.832	1.000	1.000	**0.001**	**0.000**
	Thermophilic	**0.000**	**0.000**	**0.001**	0.573	0.195	**0.000**	1.000	1.000	**0.000**
	Mesophilic	**0.000**	**0.000**	1.000	1.000	1.000	**0.000**	1.000	**0.000**	**0.000**
Oxygen requirement	Aerobic	**0.000**	**0.000**	**0.007**	1.000	1.000	0.973	1.000	1.000	**0.000**
	Anaerobic	**0.000**	**0.000**	**0.000**	0.724	**0.000**	**0.000**	1.000	1.000	**0.000**
	Facultative	**0.000**	**0.016**	**0.000**	**0.000**	**0.000**	**0.000**	1.000	**0.000**	**0.000**

B.

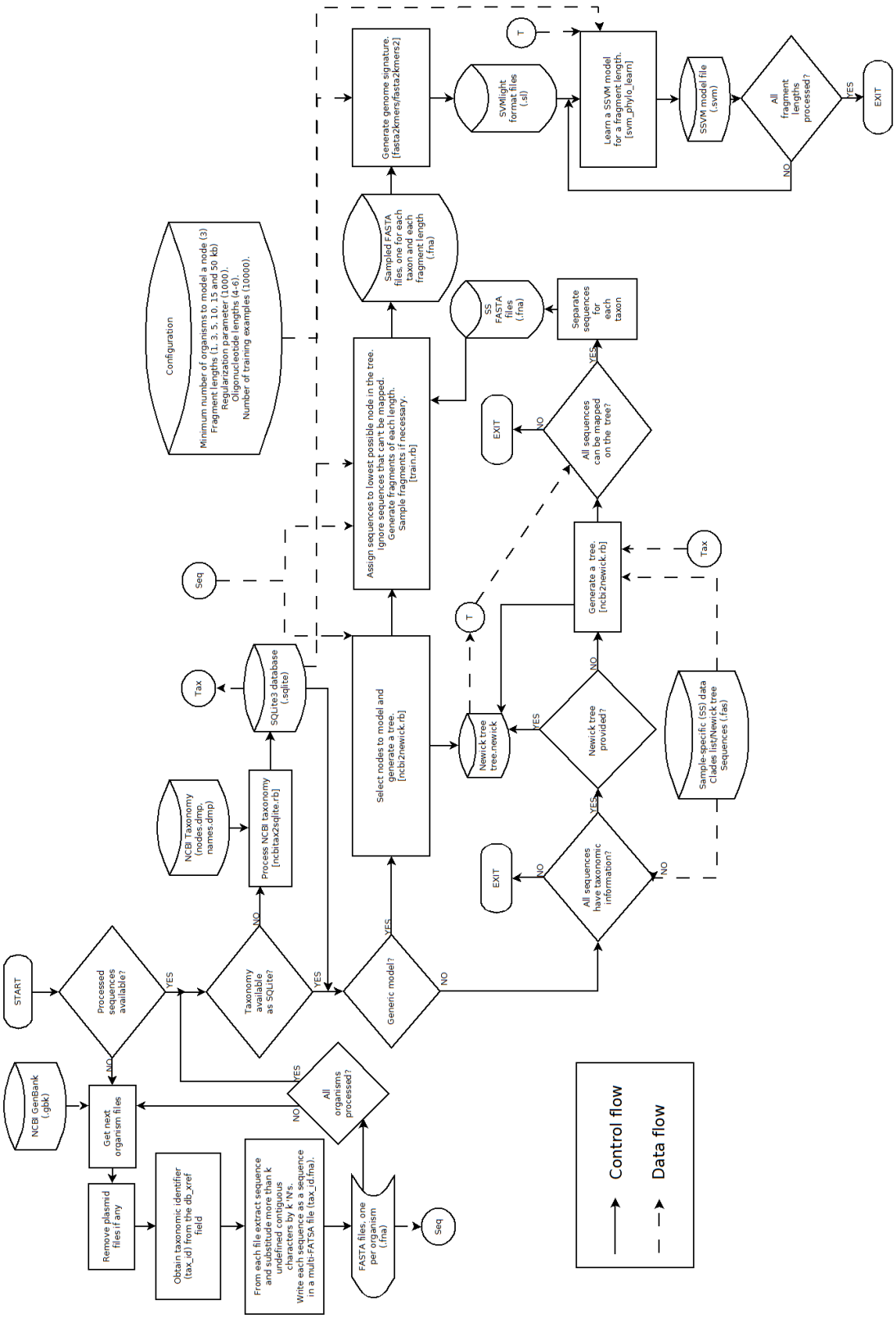
Attribute	Group	Specific l4n1	Rand l4n1	Eucl PCA l4n1	Delta 50kb l4n1	Eucl l6n1	CVTree l6n5,4	CompSpec l10r2	FFP l10ry	GBDP
Phylum	Proteobacteria	**0.004**	0.983	**0.008**	**0.000**	1.000	**0.000**	1.000	1.000	**0.000**
	Firmicutes	**0.000**	1.000	0.719	0.195	0.992	0.404	1.000	1.000	**0.000**
	Actinobacteria	**0.000**	**0.036**	**0.000**	0.245	**0.009**	**0.000**	**0.028**	**0.000**	**0.000**
	Euryarchaeota	0.052	0.925	0.196	0.082	0.650	**0.004**	0.555	0.561	**0.000**
GC-content	<=30%	**0.000**	0.965	**0.000**	**0.000**	**0.021**	**0.000**	**0.000**	**0.000**	**0.000**
	>30%-<=50%	0.689	1.000	0.999	0.975	**0.000**	**0.000**	1.000	1.000	**0.000**
	>50%-<=70%	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	1.000	1.000	**0.000**
Habitat	Aquatic	**0.001**	**0.006**	0.836	0.509	0.338	0.996	1.000	**0.000**	**0.000**
	Terrestrial	**0.005**	0.888	0.091	0.784	0.091	**0.000**	0.127	0.997	**0.000**
	Multiple	**0.000**	0.069	0.817	**0.048**	0.327	**0.000**	0.752	**0.000**	**0.000**
	Host-associated	0.555	0.984	0.468	**0.002**	**0.000**	**0.000**	0.981	0.981	**0.000**
	Specialized	**0.000**	0.497	0.959	0.940	0.992	0.990	1.000	0.989	**0.000**
Temperature range	Hyperthermophilic	**0.000**	1.000	0.518	0.300	0.628	**0.035**	0.912	**0.000**	**0.000**
	Thermophilic	**0.007**	0.986	0.668	0.153	0.821	**0.000**	0.825	**0.000**	**0.000**
	Mesophilic	0.990	0.952	1.000	**0.000**	**0.009**	**0.000**	1.000	1.000	**0.000**
Oxygen requirement	Aerobic	**0.000**	**0.000**	**0.000**	**0.001**	**0.009**	**0.000**	**0.000**	1.000	**0.000**
	Anaerobic	**0.000**	0.086	**0.000**	0.091	**0.000**	**0.000**	**0.000**	0.130	**0.000**
	Facultative	0.678	1.000	0.474	**0.005**	0.996	**0.000**	1.000	1.000	**0.000**

Supplementary Table 7. Cophenetic correlation coefficient and quartet distance before (CPCC, QD) and after (CPCC_PCA, QD_PCA) principal component analysis using the l6n1 signature. The dimension and variance columns show number of dimensions and variance retained respectively. No significant improvement was observed after applying PCA either for the CPCC or the QD (P>0.25, one-sided Wilcoxon rank sum test).

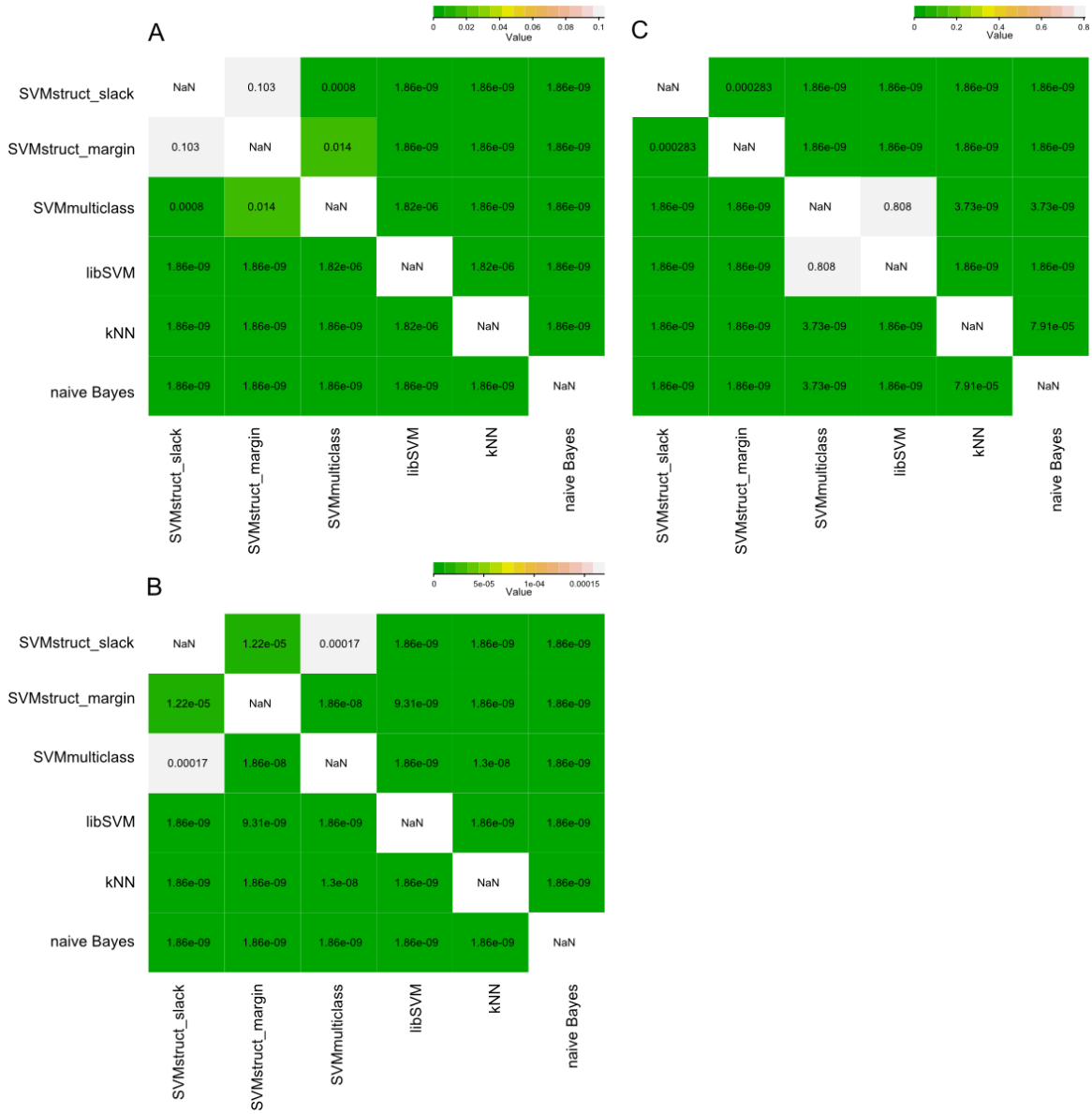
Attribute	Group	CPCC	CPCC_PCA	QD	QD_PCA	Dimension	Variance (%)
hPhylum	Proteobacteria	0.42	0.42	0.47	0.43	114	98.51
	Firmicutes	0.54	0.54	0.32	0.29	82	99.55
	Actinobacteria	0.45	0.45	0.53	0.50	63	99.83
	Euryarchaeota	0.44	0.45	0.47	0.44	48	99.95
GC-content	<=30%	0.26	0.30	0.44	0.41	54	99.86
	>30%-<=50%	0.35	0.36	0.50	0.50	137	98.47
	>50%-<=70%	0.44	0.50	0.49	0.43	122	98.51
Habitat	Aquatic	0.39	0.40	0.50	0.50	114	99.49
	Terrestrial	0.43	0.47	0.37	0.31	56	99.93
	Multiple	0.38	0.39	0.46	0.45	103	99.07
	Host-associated	0.14	0.18	0.48	0.49	114	99.05
	Specialized	0.19	0.19	0.57	0.59	93	99.82
Temperature range	Hyperthermophilic	0.44	0.39	0.46	0.42	38	99.96
	Thermophilic	0.20	0.27	0.61	0.60	61	99.91
	Mesophilic	0.24	0.25	0.48	0.50	137	97.81
Oxygen requirement	Aerobic	0.32	0.35	0.57	0.56	112	98.77
	Anaerobic	0.21	0.23	0.54	0.54	121	99.42
	Facultative	0.47	0.50	0.35	0.27	108	99.07
	AVERAGE	0.35	0.37	0.48	0.46	93.17	99.28

6.2 SUPPLEMENTARY FIGURES

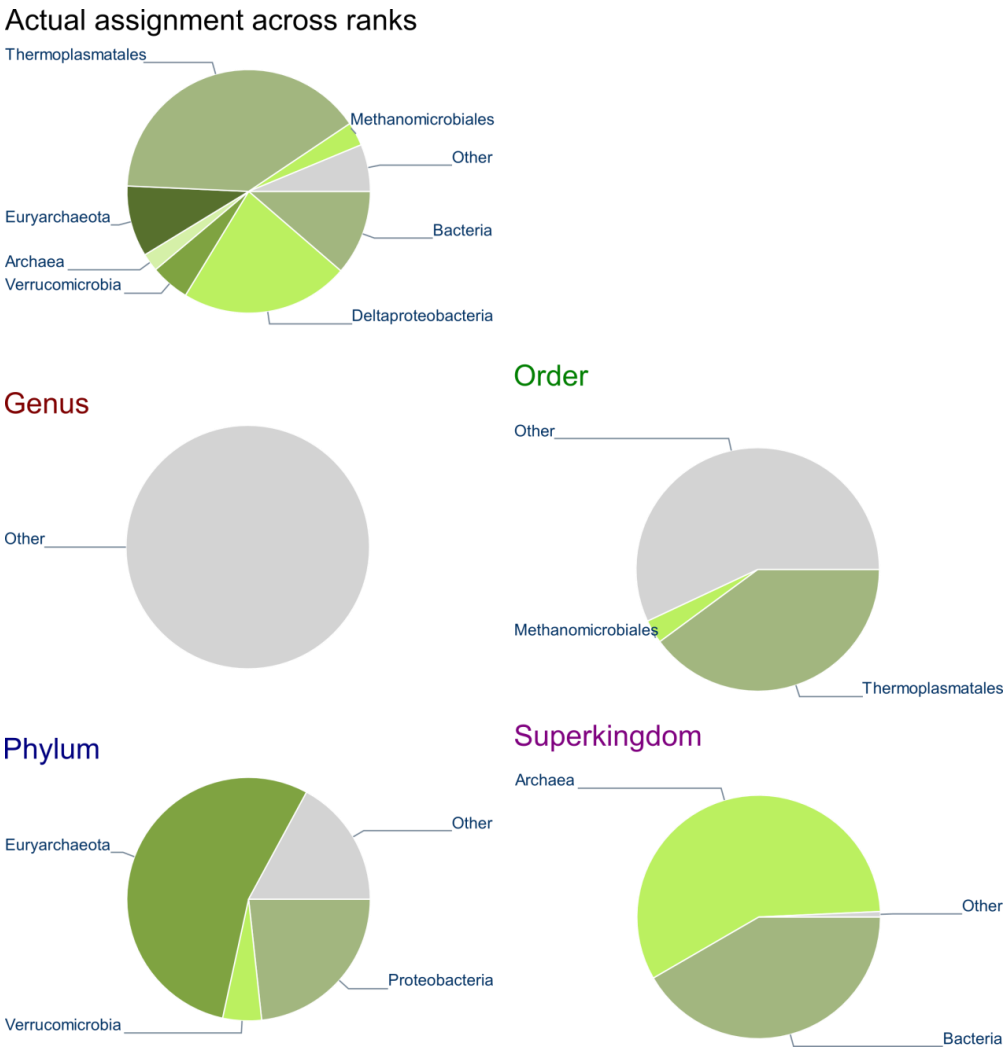
Supplementary Figure 1. The flow diagram of the PhyloPythiaS training phase.



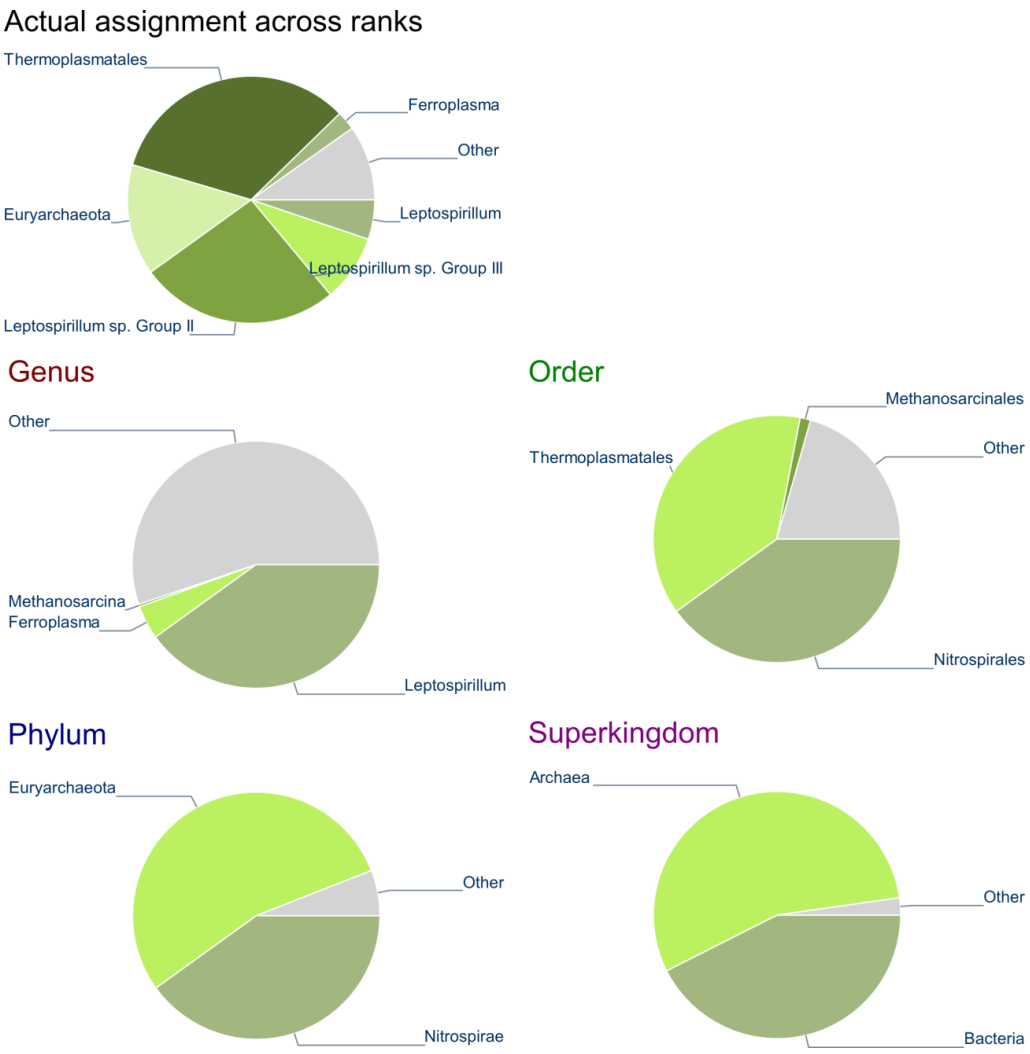
Supplementary Figure 2. Pair-wise Wilcoxon paired rank-sum test P-values for 30 folds (10 runs of 3-fold cross-validation).



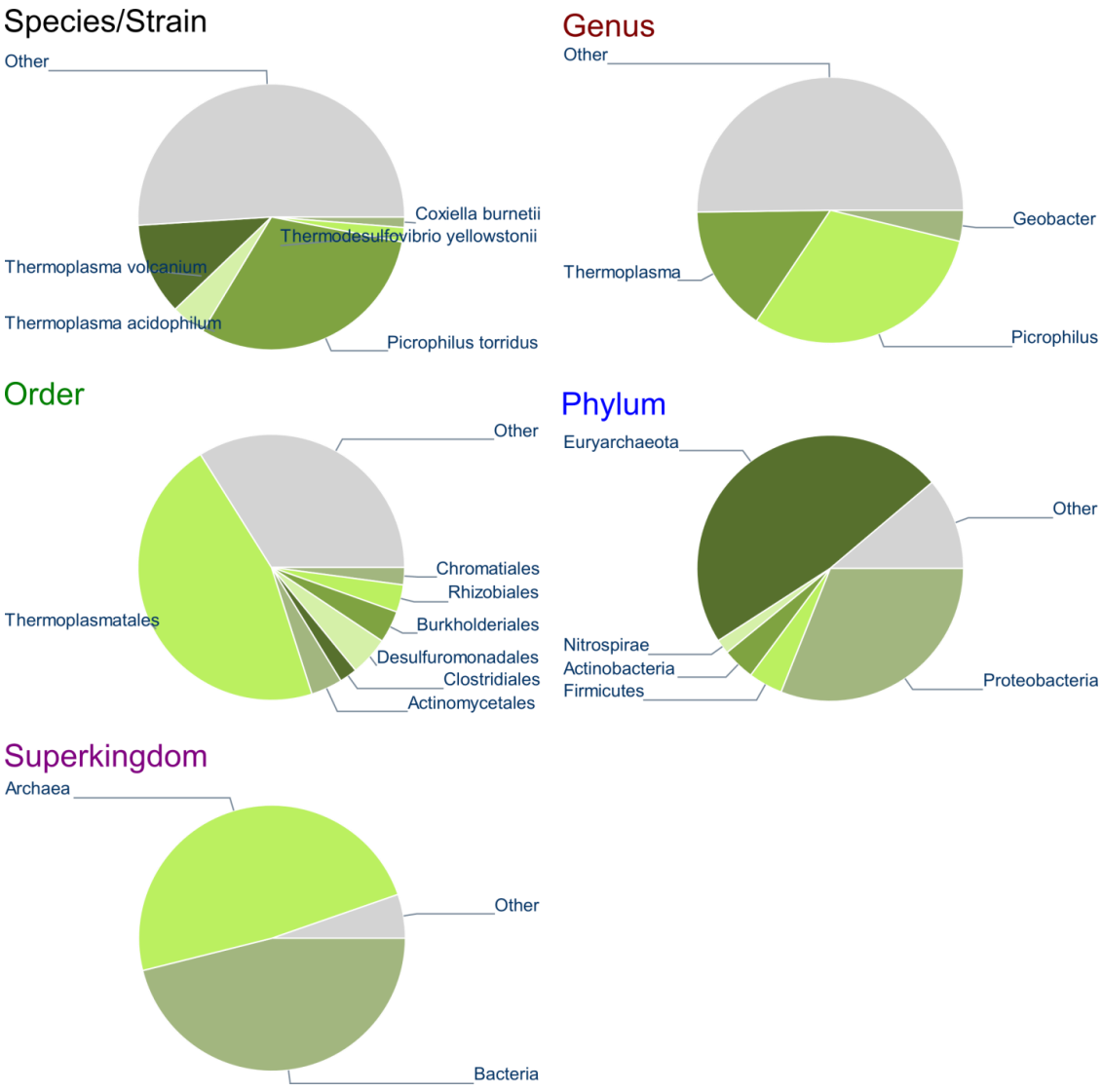
Supplementary Figure 3. Assignments for the AMD metagenome scaffolds at different taxonomic ranks by the PhyloPythiaS generic model. This model does not assign sequences to any of the genus level clades. This is expected behavior as none of the genera (Leptospirillum and Ferroplasma) were present in the generic model. The existence of Deltaproteobacteria (in Actual and Proteobacteria in Phylum) has been previously reported (Bond, Smriga, and Banfield 2000) and is due to the provisional assignment of Leptospirillium to delta subdivision (Bock and Wagner 2006).



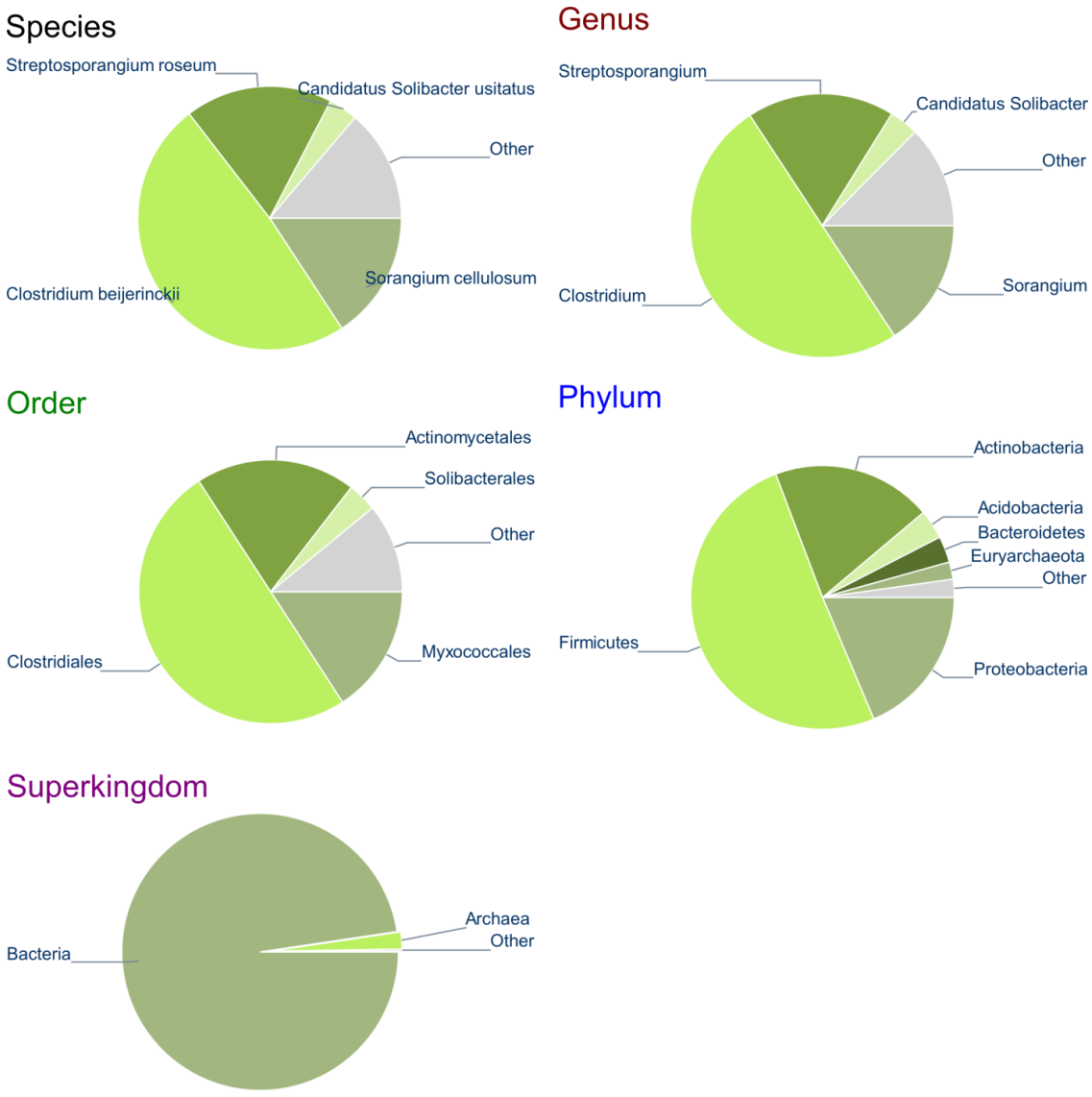
Supplementary Figure 4. Assignments for the AMD metagenome scaffolds at different taxonomic ranks by PhyloPythiaS sample-specific model. Sample specific data (approximately 100 kb from each of the four strains) from the two genera (Leptospirillum and Ferroplasma) was used.



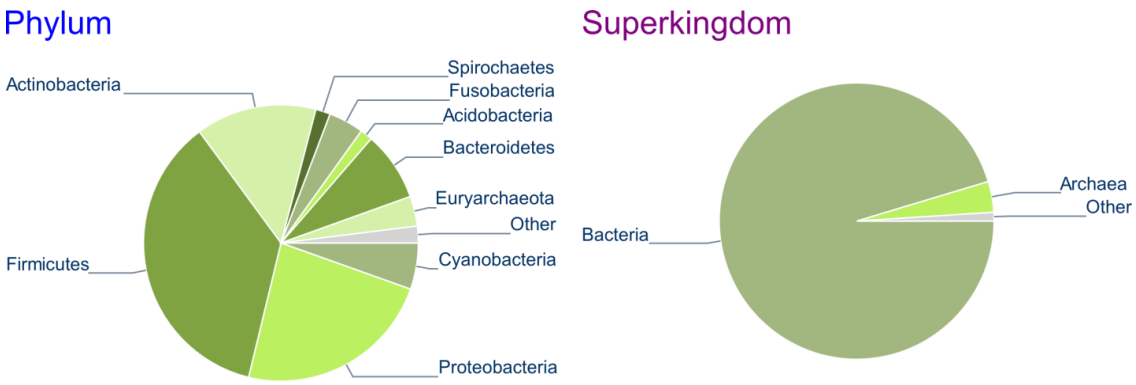
Supplementary Figure 5. Assignments for the AMD metagenome scaffolds at different taxonomic ranks by best BLASTN hit with e-value cut-off of 0.1. The blast database used same genomes used for creating PhyloPythiaS generic model, i.e. all 1076 complete genomes available from NCBI as of April 2010.



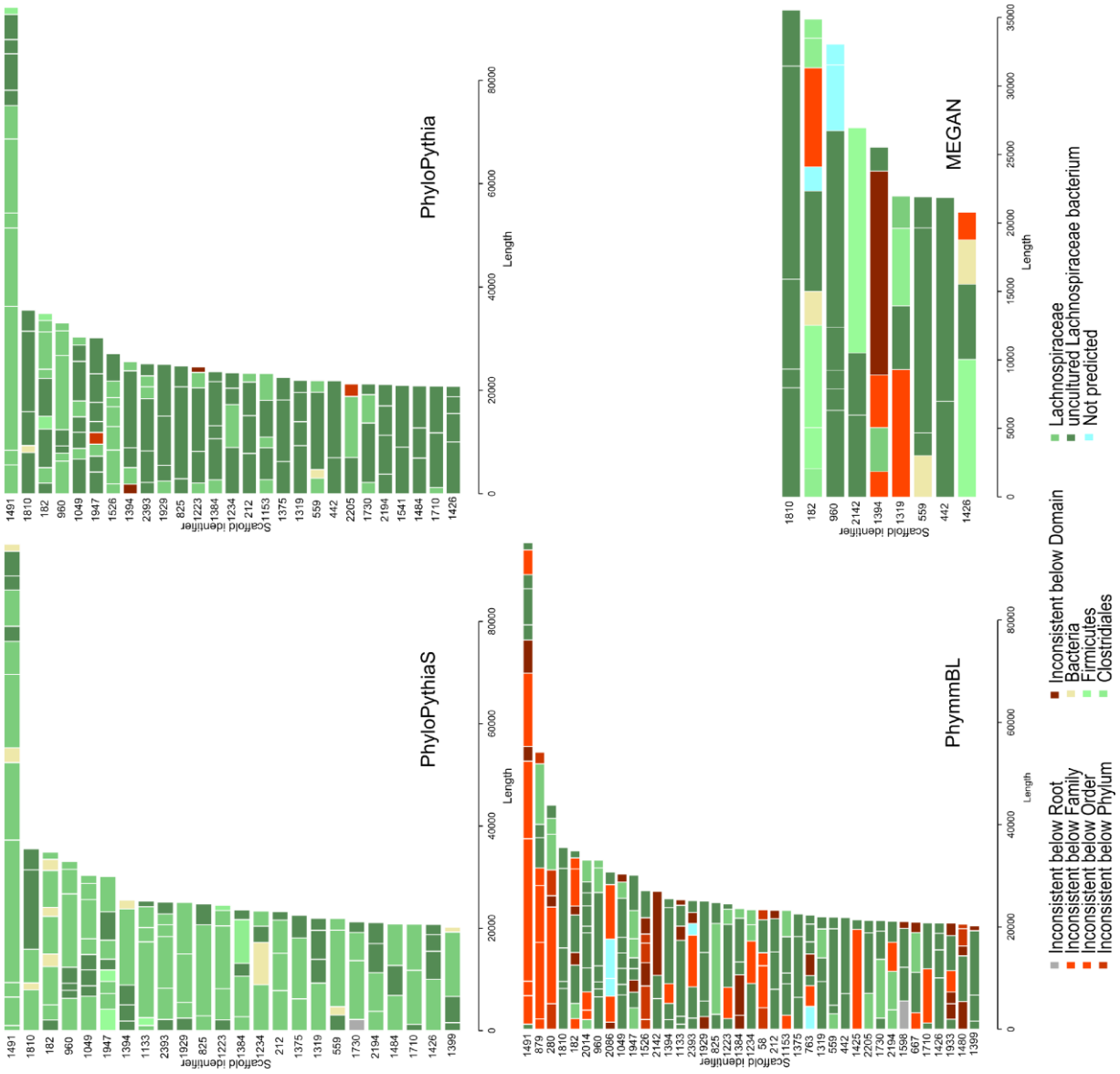
Supplementary Figure 6. Assignments for the AMD metagenome scaffolds at different taxonomic ranks by the NBC webserver. Default N-mer length of 15 with Bacteria/Archaea genomes were used. The webserver was accessed at http://nbc.ece.drexel.edu/ in April 2011.



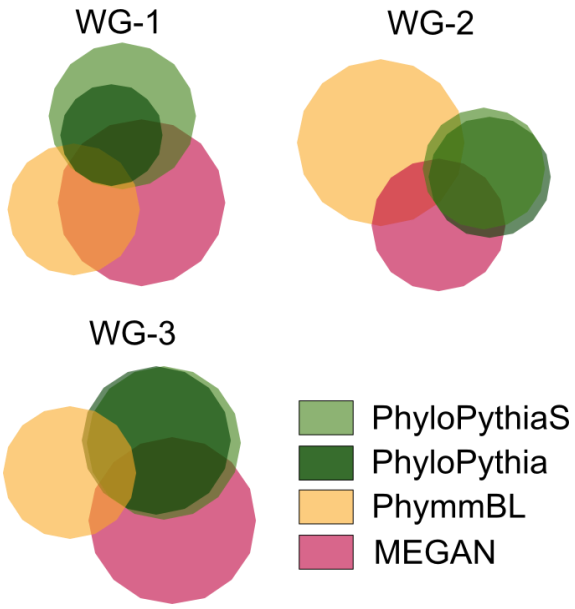
Supplementary Figure 7. Assignments for the AMD metagenome (scaffolds fragmented at 500 bp) at different taxonomic ranks by the NBC webserver. To check for the possible effect of test sequence length on the taxonomic assignment of the AMD metagenome using the NBC webserver, we created fragments of length 500 bp from the scaffolds and obtained their assignments. Default N-mer length of 15 and Bacteria/Archaea genomes were used. Bacteria were overestimated while underestimating the Archaea. The NBC webserver was accessed at http://nbc.ece.drexel.edu/ in May 2011.



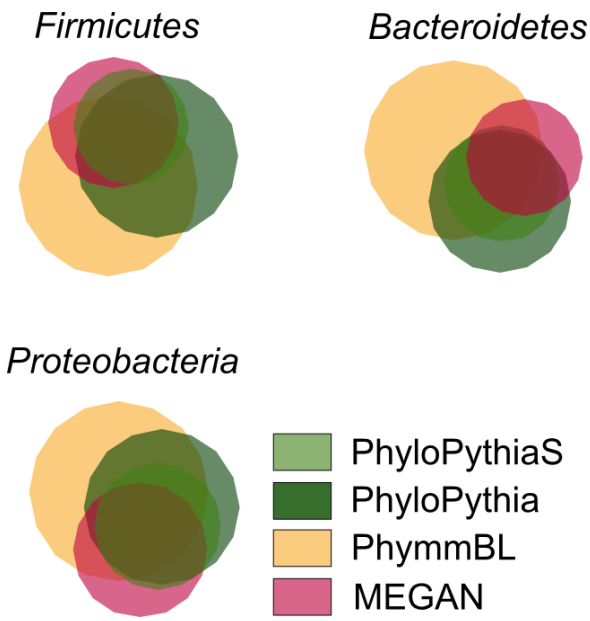
Supplementary Figure 8. Scaffold-contig visualization of different binning methods for the WG-2 population from the TW sample. Every horizontal bar represents a scaffold and its constituent contigs. Every contig is color coded to represent its consistency with respect to the scaffold assignment. Only scaffolds >=20 kb in length are shown for clarity.



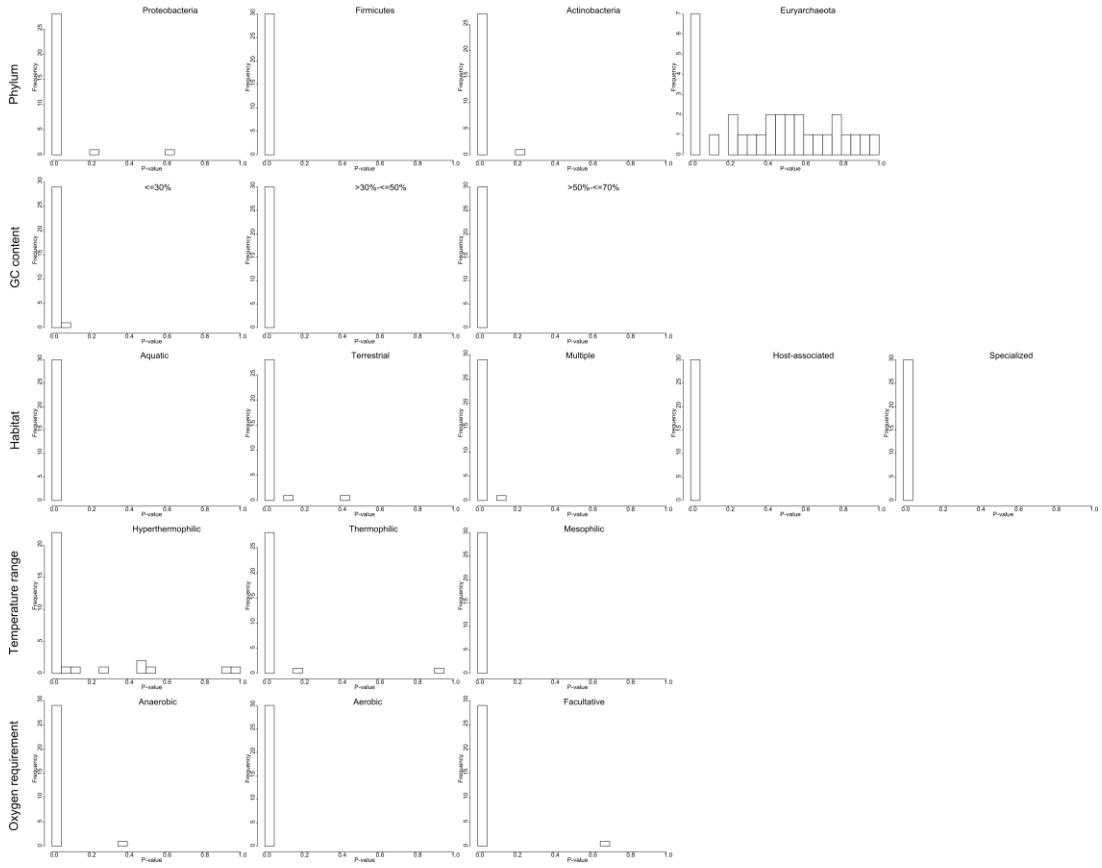
Supplementary Figure 9. Overlap between predictions of different methods on the TW sample for the three uncultured populations. The overlaps are represented as area proportional Euler diagrams. Only exact predictions were taken into account for each population. The areas correspond to the predictions of the methods on the union of contigs predicted as a particular clade by at least one method. As it can be seen, PhyloPythiaS and PhyloPythia have large overlaps for all populations.



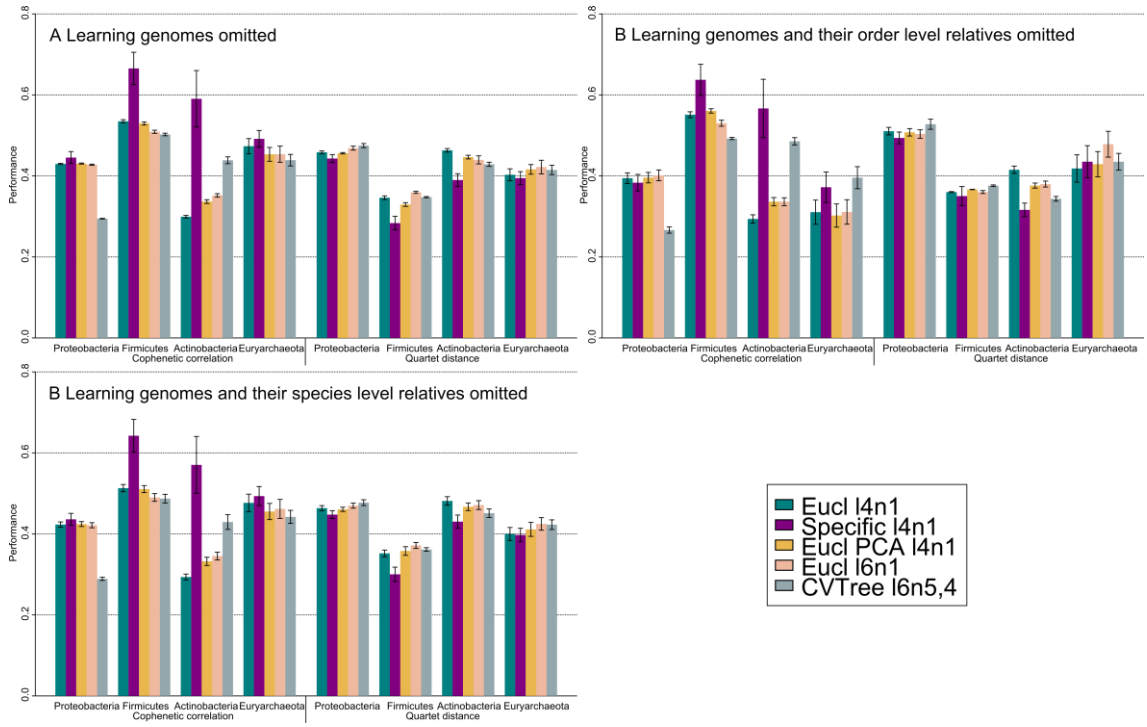
Supplementary Figure 10. Overlap between predictions of different methods on TW sample for dominant phyla. The overlaps are represented as area proportional Euler diagrams. The areas correspond to the predictions of the methods on the union of contigs predicted as a particular clade by at least one method. All the predictions were mapped to its corresponding phyla. As it can be seen, PhyloPythiaS, PhyloPythia and MEGAN have large overlaps for all three phyla.



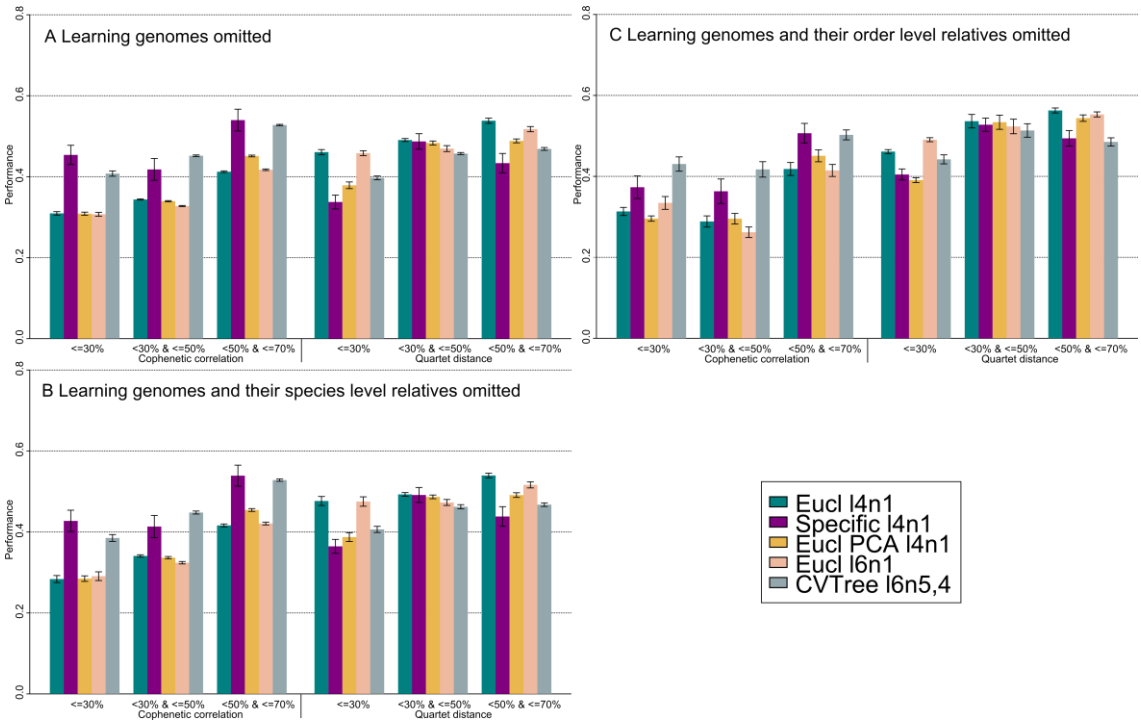
Supplementary Figure 11. Histograms of P-values computed using the Hotelling-Williams test for dependent correlation coefficients that share a variable. Here the shared variable is phenetic distances derived from taxonomy and change in correlation is considered with respect to the baseline correlation. Each box shows histogram of 30 P-values for a group.



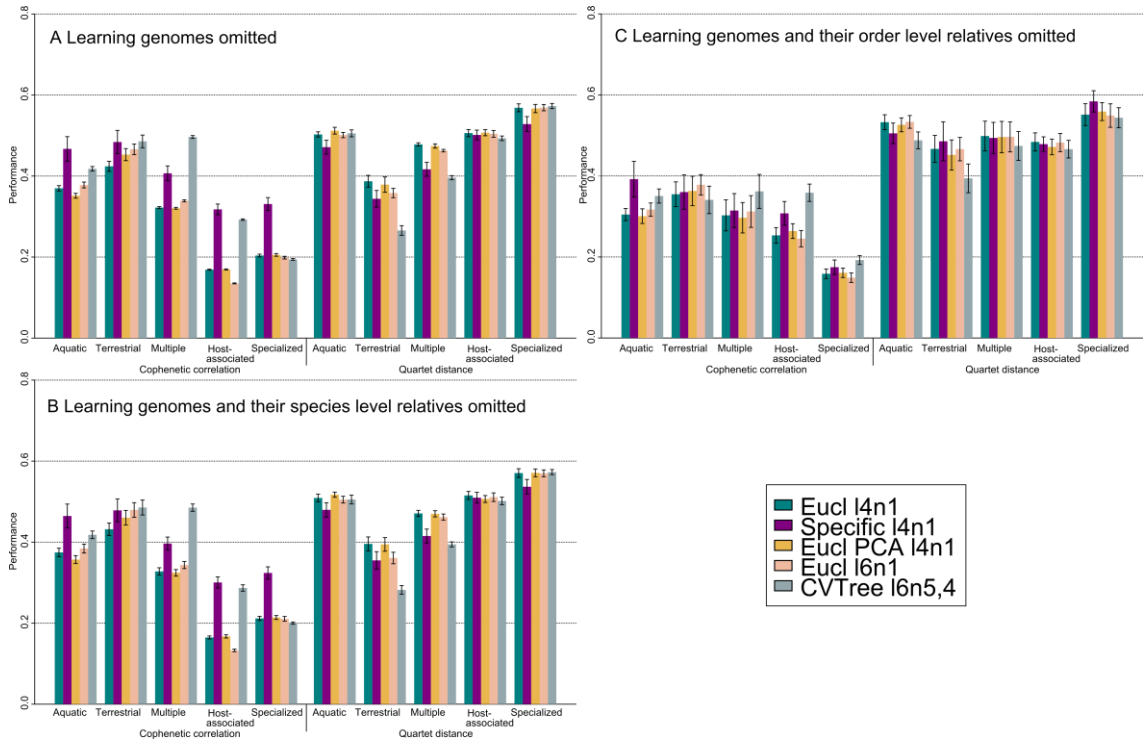
Supplementary Figure 12. Performance of the metrics on four phylogenetic groups after removing genomes used for learning and their species and order level relatives.



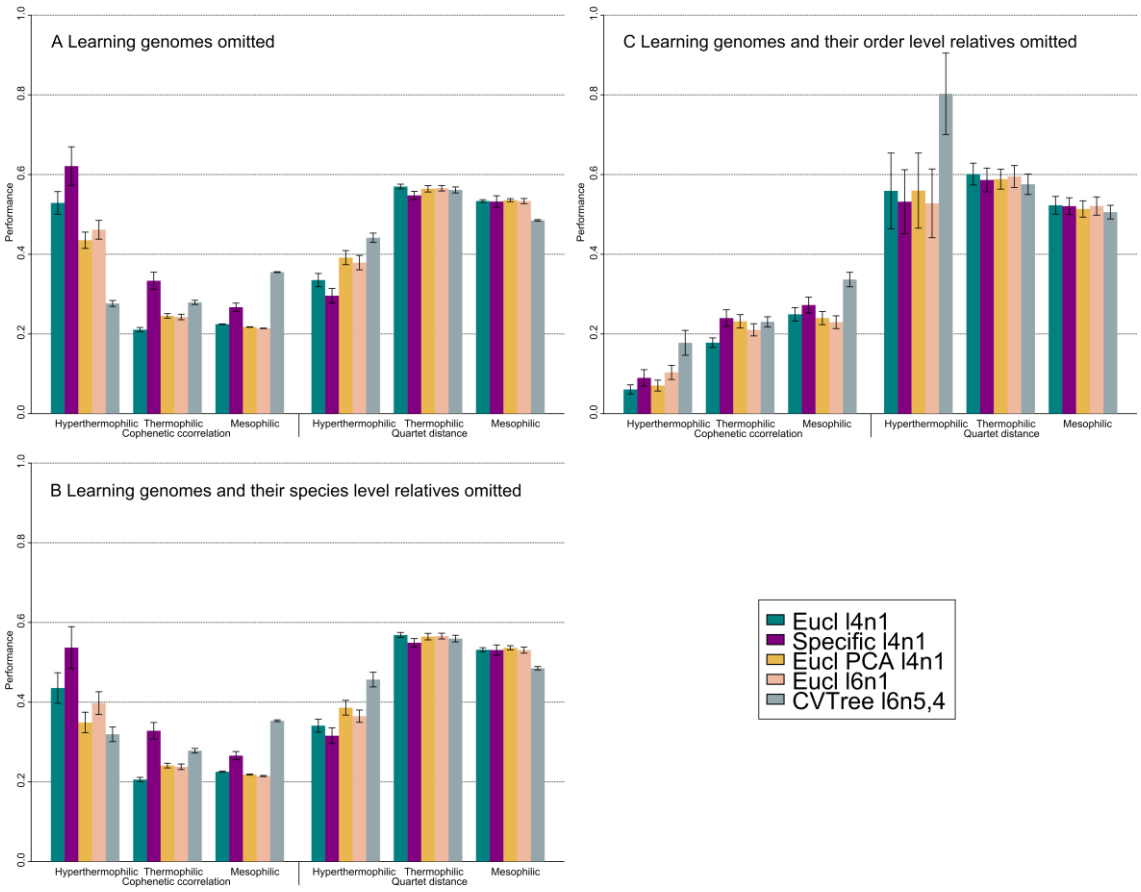
Supplementary Figure 13. Performance of the metrics on the GC content groups after removing genomes related to the learning genomes at species and order ranks.



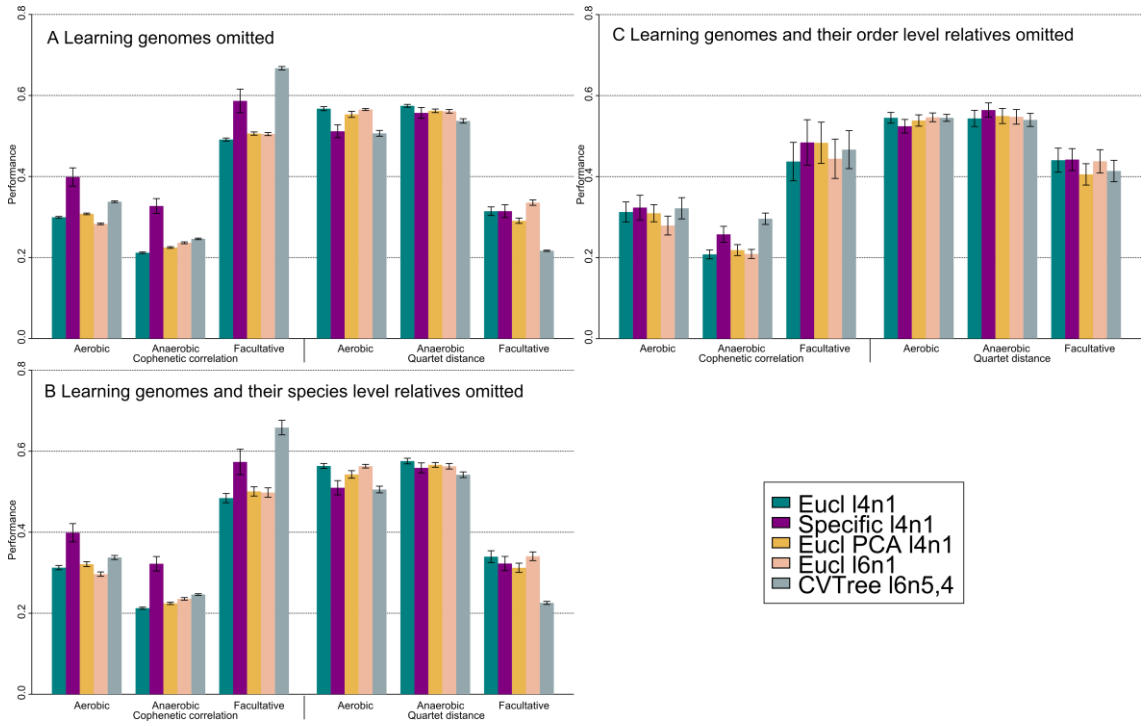
Supplementary Figure 14. Performance of the metrics on the habitat groups after removing genomes related to the learning genomes at species and order ranks.



Supplementary Figure 15. Performance of the metrics on the temperature range groups after removing genomes related to the learning genomes at species and order ranks.



Supplementary Figure 16. Performance of the metrics on the Oxygen requirement groups after removing genomes related to the learning genomes at species and order ranks.



BIBLIOGRAPHY

Abe T, Sugawara H, Kinouchi M, Kanaya S, & Ikemura T. 2005. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Research. 12(5) pp. 281–290.

Adams N. 2010. Dataset Shift in Machine Learning. Journal of the Royal Statistical Society Series a-Statistics in Society. 173 pp. 274.

Albrecht-Buehler G. 2006. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proceedings of the National Academy of Sciences of the United States of America. 103(47) pp. 17828–17833.

Almagor H. 1983. A Markov analysis of DNA sequences. Journal of Theoretical Biology. 104(4) pp. 633–645.

Almeida JS, Carrico JA, Maretzek A, Noble PA, & Fletcher M. 2001. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics. 17(5) pp. 429–437.

Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology. 215(3) pp. 403–410.

Altun Y, Tsochantaridis I, & Hofmann T. 2003. Hidden Markov Support Vector Machines. 20th International Conference on Machine Learning.

Andersson SG, & Sharp PM. 1996. Codon usage and base composition in Rickettsia prowazekii. Journal of Molecular Evolution. 42(5) pp. 525–536.

Baldi P, & Brunak S. 2001. *Bioinformatics: The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning)*. The MIT Press.

Bapteste E et al. 2009. Prokaryotic evolution and the tree of life are two different things. Biology Direct. 4 pp. -.

Basak S, Mandal S, & Ghosh TC. 2005. Correlations between genomic GC levels and optimal growth temperatures: some comments. Biochemical and Biophysical Research Communications. 327(4) pp. 969–970.

Bentley SD, & Parkhill J. 2004. Comparative genomic structure of prokaryotes. Annual Review of Genetics. 38 pp. 771–792.

Beyer KS, Goldstein J, Ramakrishnan R, & Shaft U. 1999. When Is Nearest Neighbor'' Meaningful? In: ICDT '99 Proceedings of the 7th International Conference on Database Theory. Springer-Verlag London, UK pp. 217–235.

Blaisdell BE. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proceedings of the National Academy of Sciences of the United States of America. 83(14) pp. 5155–5159.

Bohlin J, Skjerve E, & Ussery DW. 2009. Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. BMC Genomics. 10 pp. -.

Bohlin J, Skjerve E, & Ussery DW. 2008. Investigations of oligonucleotide usage variance within and between prokaryotes. PLoS Computational Biology. 4(4) pp. -.

Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Donsvik T, Skjerve E, & Ussery DW. 2010. Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics. 11 pp. -.

Boser B, Guyon I, & Vapnik VN. 1992. A training algorithm for optimal margin classifiers. In: Haussler, D, editor. ACM Press pp. 144–152.

Botzman M, & Margalit H. 2011. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biology. 12(10) pp. R109.

Brady A, & Salzberg SL. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods. 6(9) pp. 673–676.

Burge C, Campbell AM, & Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proceedings of the National Academy of Sciences of the United States of America. 89(4) pp. 1358–1362.

Cameron M, Bernstein Y, & Williams HE. 2007. Clustered sequence representation for fast homology search. Journal of Computational Biology. 14(5) pp. 594–614.

Campbell A, Mrazek J, & Karlin S. 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America. 96(16) pp. 9184–9189.

Casjens S. 1998. The diverse and dynamic structure of bacterial genomes. Annual Review of Genetics. 32 pp. 339–377.

Cesa-Bianchi N, Gentile C, & Zaniboni L. 2006. Incremental algorithms for hierarchical classification. Journal of Machine Learning Research. 7 pp. 31–54.

Chargaff E. 1950. Chemical Specificity of Nucleic Acids and Mechanism of Their Enzymatic Degradation. Experientia. 6(6) pp. 201–209.

Chargaff E. 1951. Structure and function of nucleic acids as cell constituents. Federation Proceedings. 10(3) pp. 654–659.

Chen SL, Lee W, Hottes AK, Shapiro L, & McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences of the United States of America. 101(10) pp. 3480–3485.

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, & Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311(5765) pp. 1283–1287.

Coenye T, & Vandamme P. 2004. Use of the genomic signature in bacterial classification and identification. Systematic and Applied Microbiology. 27(2) pp. 175–185.

Cohan FM, & Koeppel AF. 2008. The origins of ecological diversity in prokaryotes. Current Biology. 18(21) pp. R1024–34.

Cortes C, & Vapnik V. 1995. Support-Vector Networks. Machine Learning. 20(3) pp. 273–297.

Cover TM, & Hart PE. 1967. Nearest Neighbor Pattern Classification. Ieee Transactions on Information Theory. 13(1) pp. 21–+.

Crammer K, & Singer Y. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research. 2(2) pp. 265–292.

Crick FH, Barnett L, Brenner S, & Watts-Tobin RJ. 1961. General nature of the genetic code for proteins. Nature. 192 pp. 1227–1232.

Curtis TP, Sloan WT, & Scannell JW. 2002. Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences of the United States of America. 99(16) pp. 10494–10499.

Darwin C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London : John Murray.

Delcher AL, Phillippy A, Carlton J, & Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30(11) pp. 2478–2483.

Delsuc F, Brinkmann H, & Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6(5) pp. 361–375.

DeSantis TZ et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology. 72(7) pp. 5069–5072.

Deschavanne PJ, Giron A, Vilain J, Fagot G, & Fertil B. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Molecular Biology and Evolution. 16(10) pp. 1391–1399.

Diaz NN, Krause L, Goesmann A, Niehaus K, & Nattkemper TW. 2009. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. Bmc Bioinformatics. 10 pp. 56.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Yelton AP, & Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biology. 10(8).

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284(5423) pp. 2124–2129.

Doolittle WF. 2000. Uprooting the tree of life. Scientific American. 282(2) pp. 90–95.

Dröge J, & McHardy AC. 2012. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Briefings in Bioinformatics.

Dufraigne C, Fertil B, Lespinats S, Giron A, & Deschavanne P. 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 33(1) pp. e6.

Evgeniou T, Poggio T, Pontil M, & Verri A. 2002. Regularization and statistical learning theory for data analysis. Computational Statistics & Data Analysis. 38(4) pp. 421–432.

Evgeniou T, & Pontil M. 2004. Regularized multi--task learning. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.

Farris JS. 1969. On the Cophenetic Correlation Coefficient. Systematic Zoology. 18(3) pp. 279–285.

Foerstner KU, Von Mering C, Hooper SD, & Bork P. 2005. Environments shape the nucleotide composition of genomes. EMBO Rep. 6(12) pp. 1208–1213.

Forsdyke DR, & Mortimer JR. 2000. Chargaff's legacy. Gene. 261(1) pp. 127–137.

Fox G et al. 1980. The phylogeny of prokaryotes. Science. 209(4455) pp. 457–463.

Garcia Martin H et al. 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nature Biotechnology. 24(10) pp. 1263–1269.

Garrity G. 2005. *Bergey's manual of systematic bacteriology: The proteobacteria. Introductory essays, Part 1*. 2, illustr. Springer.

Gerlach W, Junemann S, Tille F, Goesmann A, & Stoye J. 2009. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. Bmc Bioinformatics. 10 pp. 430.

Gill SR et al. 2006. Metagenomic analysis of the human distal gut microbiome. Science. 312(5778) pp. 1355–1359.

Glossary(Genome). Genome Glossary. URL:http://www.ornl.gov/sci/techresources/Human_Genome/glossary/ Program, USD of EG, editor. 2012.

Glossary(NCBI). 2002. The NCBI Handbook [Internet]. URL: http://www.ncbi.nlm.nih.gov/books/NBK21106/ McEntyre J, OJ, editor. Glosarry.

Glossary(Systematics). Palaeos - Systematics, Taxonomy, and Phylogeny: Glossary. URL: http://palaeos.com/systematics/glossary.html.

Goto N, Prins P, Nakao M, Bonnal R, Aerts J, & Katayama T. 2010. BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics. 26(20) pp. 2617–2619.

Grantham R. 1980. Workings of the Genetic-Code. Trends in Biochemical Sciences. 5(12) pp. 327–331.

Grantham R, Gautier C, Gouy M, Mercier R, & Pave A. 1980. Codon Catalog Usage and the Genome Hypothesis. Nucleic Acids Research. 8(1) pp. R49–R62.

Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews. 68(4) pp. 669–685.

Handelsman J, Rondon MR, Brady SF, Clardy J, & Goodman RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chemistry and Biology. 5(10) pp. R245–R249.

Hansen N, Muller SD, & Koumoutsakos P. 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evolutionary Computation. 11(1) pp. 1–18.

Hao B, & Qi J. 2003. Prokaryote phylogeny without sequence alignment: From avoidance signature to composition distance. Proceedings of the 2003 Ieee Bioinformatics Conference. pp. 375–384.

Hastie T, Tibshirani R, & Friedman JH. 2009. *The elements of statistical learning : data mining, inference, and prediction*. 2nd ed. Springer: New York, NY.

Hess M et al. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 331(6016) pp. 463–467.

Hirschman L et al. 2008. Habitat-Lite: A GSC case study based on free text terms for environmental metadata. Omics-a Journal of Integrative Biology. 12(2) pp. 129–136.

Hsu C-W, & Lin C-J. 2002. A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on. 13(2) pp. 415–425.

Hugenholtz P. 2002. Exploring prokaryotic diversity in the genomic era. Genome Biol. 3(2).

Hugenholtz P, Goebel BM, & Pace NR. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology. 180(18) pp. 4765–4774.

Hurst LD, & Merchant AR. 2001. High guanineâ€“cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proceedings of the Royal Society of London. Series B: Biological Sciences. 268(1466) pp. 493–497.

Huson DH, Auch AF, Qi J, & Schuster SC. 2007. MEGAN analysis of metagenomic data. Genome Research. 17(3) pp. 377–386.

Huson DH, Richter DC, Mitra S, Auch AF, & Schuster SC. 2009. Methods for comparative metagenomics. Bmc Bioinformatics. 10 Suppl 1 pp. S12.

Huson DH, & Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Systematic biology. 61(6) pp. 1061–7.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Molecular Biology and Evolution. 2(1) pp. 13–34.

Jeffrey HJ. 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18(8) pp. 2163–2170.

Jensen LJ et al. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 37(Database issue) pp. D412–6.

Joachims T, Finley T, & Yu C-N. 2009. Cutting-plane training of structural SVMs. Machine Learning. 77(1) pp. 27–59.

Josse J, Kaiser AD, & Kornberg A. 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. Journal of Biological Chemistry. 236 pp. 864–875.

Kahn SD. 2011. On the future of genomic data. Science. 331(6018) pp. 728–729.

Kampfer P, & Glaeser SP. 2012. Prokaryotic taxonomy in the sequencing era - the polyphasic approach revisited. Environ Microbiol. 14(2) pp. 291–317.

Karlin S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Current Opinion in Microbiology. 1(5) pp. 598–610.

Karlin S. 1994. Statistical studies of biomolecular sequences: score-based methods. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 344(1310) pp. 391–402.

Karlin S, & Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics. 11(7) pp. 283–290.

Karlin S, Campbell AM, & Mrazek J. 1998. Comparative DNA analysis across diverse genomes. Annual Review of Genetics. 32 pp. 185–225.

Karlin S, & Cardon LR. 1994. Computational DNA Sequence Analysis. Annual Review of Microbiology. 48(1) pp. 619–654.

Karlin S, & Mrazek J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. Journal of Bacteriology. 182(18) pp. 5238–5250.

Karlin S, Mrazek J, & Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. Journal of Bacteriology. 179(12) pp. 3899–3913.

Kestler HA, Muller A, Kraus JM, Buchholz M, Gress TM, Liu H, Kane DW, Zeeberg BR, & Weinstein JN. 2008. VennMaster: area-proportional Euler diagrams for functional GO analysis of microarrays. Bmc Bioinformatics. 9 pp. 67.

Kirzhner V, Korol A, Bolshoy A, & Nevo E. 2002. Compositional spectrum - revealing patterns for genomic sequence characterization and comparison. Physica A. 312(3-4) pp. 447–457.

Kirzhner V, Paz A, Volkovich Z, Nevo E, & Korol A. 2007a. Different Clustering of Genomes Across Life Using the A-T-C-G and Degenerate R-Y Alphabets: Early and Late Signaling on Genome Evolution? Journal of Molecular Evolution. 64(4) pp. 448–456.

Kirzhner V, Paz A, Volkovich Z, Nevo E, & Korol A. 2007b. Different clustering of genomes across life using the A-T-C-G and degenerate R-Y alphabets: early and late signaling on genome evolution? Journal of Molecular Evolution. 64(4) pp. 448–456.

Kohavi R, & Provost F. 1998. Glossary of Terms. Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. Machine Learning. 30(2/3) pp. 271–274.

Koonin E V. 2012. Does the central dogma still stand? Biology Direct. 7(1) pp. 27.

Koonin E V. 2005. Orthologs, paralogs, and evolutionary genomics. Annual review of genetics. 39 pp. 309–38.

Koonin E V, Makarova KS, & Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annual Review of Microbiology. 55 pp. 709–742.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, & Stoye J. 2008. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 36(7) pp. 2230–2239.

Kryukov K, Sumiyama K, Ikeo K, Gojobori T, & Saitou N. 2012. A new database (GCD) on genome composition for eukaryote and prokaryote genome sequences and their initial analyses. Genome Biol Evol. 4(4) pp. 501–512.

Larkin MA et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics. 23(21) pp. 2947–2948.

Leslie C, Eskin E, & Noble WS. 2002. The spectrum kernel: a string kernel for SVM protein classification. Pacific Symposium on Biocomputing. pp. 564–575.

Levene PA. 1919. The structure of yeast nucleic acid. IV. Ammonia hydrolysis. Journal of Biological Chemistry. 40(2) pp. 415–424.

Li Q, Xu Z, & Hao B. 2010. Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. Journal of Biotechnology. 149(3) pp. 115–119.

Lobry JR, & Sueoka N. 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. 3(10) pp. RESEARCH0058.

Mahalanobis PC. 1936. On the generalised distance in statistics. In: Proceedings National Institute of Science, India.Vol. 2 pp. 49–55.

Mann S, & Chen YP. 2010. Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics. 95(1) pp. 7–15.

Marashi SA, & Ghalanbor Z. 2004. Correlations between genomic GC levels and optimal growth temperatures are not “robust”. Biochemical and Biophysical Research Communications. 325(2) pp. 381–383.

Mavromatis K et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods. 4(6) pp. 495–500.

McEwan CEA, Gatherer D, & McEwan NR. 1998. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas. 128(2) pp. 173–178.

McHardy AC, Garcia-Martin H, Tsirigos A, Hugenholtz P, & I. R. 2007. Accurate Phylogenetic Classification of Variable-length DNA fragments. 4(1) pp. 63–72.

McHardy AC, Puhler A, Kalinowski J, & Meyer F. 2004. Comparing expression level-dependent features in codon usage with protein abundance: an analysis of “predictive proteomics”. Proteomics. 4(1) pp. 46–58.

McHardy AC, & Rigoutsos I. 2007. What’s in the mix: phylogenetic classification of metagenome sequence samples. Current Opinion in Microbiology. 10(5) pp. 499–503.

McLean MJ, Wolfe KH, & Devine KM. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. Journal of Molecular Evolution. 47(6) pp. 691–696.

Von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, & Bork P. 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science. 315(5815) pp. 1126–1130.

Metagenomics(NCBI). 2006. Metagenomics: Sequences from the Environment [Internet] (US), B (MD): NC for BI, editor. 2012(12 September 2012).

Metagenomics(NRC). 2007. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press.

Metzker ML. 2010. Sequencing technologies - the next generation. Nature Reviews Genetics. 11(1) pp. 31–46.

Meyer F et al. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. Bmc Bioinformatics. 9 pp. 386.

Mitchell TM. 1997. *Machine Learning*. McGraw-Hill.

Monzoorul Haque M, Ghosh TS, Komanduri D, & Mande SS. 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics. 25(14) pp. 1722–1730.

Mooers AO, & Holmes EC. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol Evol. 15(9) pp. 365–369.

Moran NA, McCutcheon JP, & Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annual Review of Genetics. 42 pp. 165–190.

Mrazek J. 2009. Phylogenetic signals in DNA composition: limitations and prospects. Molecular Biology and Evolution. 26(5) pp. 1163–1169.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, & Bernardi G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Letters. 573(1-3) pp. 73–77.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, & Bernardi G. 2005. The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: A reply to Marashi and Ghalanbor. Biochemical and Biophysical Research Communications. 330(2) pp. 357–360.

Muto A, & Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proceedings of the National Academy of Sciences of the United States of America. 84(1) pp. 166–169.

Naya H, Romero H, Zavala A, Alvarez B, & Musto H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. Journal of Molecular Evolution. 55(3) pp. 260–264.

Needleman SB, & Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 48(3) pp. 443–453.

Van Niel CB. 1946. THE CLASSIFICATION AND NATURAL RELATIONSHIPS OF BACTERIA. Cold Spring Harb Symp Quant Biol. 11(0) pp. 285–301.

Nielsen J, Kristensen AK, Mailund T, & Pedersen CNS. 2011. A sub-cubic time algorithm for computing the quartet distance between two general trees. Algorithms for Molecular Biology. 6.

Orla-Jensen S. 1909. Die Hauptlinien des nattirlichen Bakteriensystems nebst einer Uebersicht der Garungsphenomene. Zentr. Bakt. Parasitenk. II(22) pp. 305–346.

Orr MR, & Smith TB. 1998. Ecology and speciation. Trends Ecol Evol. 13(12) pp. 502–506.

Parks D, MacDonald N, & Beiko R. 2011. Classifying short genomic fragments from novel lineages using composition and homology. Bmc Bioinformatics. 12(1) pp. 328.

Van Passel MW, Kuramae EE, Luyf AC, Bart A, & Boekhout T. 2006. The reach of the genome signature in prokaryotes. BMC Evol Biol. 6 pp. 84.

Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, & McHardy AC. 2011. Taxonomic metagenome sequence assignment with structured output models. Nat Methods. 8(3) pp. 191–192.

Patil KR, Roune L, & McHardy AC. 2012. The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. PLoS One. 7(6) pp. e38581.

Pertsemlidis A, & Fondon J. 2001. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biology. 2(10) pp. reviews2002.1 – reviews2002.10.

Pimenta E, & Gama J. 2005. A study on Error Correcting Output Codes. 2005 Portuguese Conference on Artificial Intelligence, Proceedings. pp. 218–223.

PlanetMath. PlanetMath.org.

Platt JC, Cristianini N, & Shawe-Taylor J. 2000. Large margin DAGs for multiclass classification. Advances in Neural Information Processing Systems 12. 12 pp. 547–553.

Pop M, Kosack DS, & Salzberg SL. 2004. Hierarchical scaffolding with Bambus. Genome research. 14(1) pp. 149–59.

Pope PB et al. 2010. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. Proceedings of the National Academy of Sciences of the United States of America. 107(33) pp. 14793–14798.

Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, Morrison M, & Eijsink VGH. 2012. Metagenomics of the Svalbard Reindeer Rumen Microbiome Reveals Abundance of Polysaccharide Utilization Loci. PLoS One. 7(6) pp. e38571.

Pope PB, Smith W, Denman SE, Tringe SG, Barry K, Hugenholtz P, McSweeney CS, McHardy AC, & Morrison M. 2011. Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. Science. 333(6042) pp. 646–648.

Pride DT, & Blaser MJ. 2002. Identification of Horizontally Acquired Genetic Elements in Helicobacter pylori and Other Prokaryotes Using Oligonucleotide Difference Analysis. Genome Letters. 1(1) pp. 2–15.

Pride DT, Meinersmann RJ, Wassenaar TM, & Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Research. 13(2) pp. 145–158.

Qi J, Luo H, & Hao B. 2004a. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res. 32(Web Server issue) pp. W45–7.

Qi J, Wang B, & Hao B. 2004b. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. Journal of Molecular Evolution. 58(1) pp. 1–11.

Rifkin R, & Klautau A. 2004. In defense of one-vs-all classification. Journal of Machine Learning Research. 5 pp. 101–141.

Rocha EP. 2008. Evolutionary patterns in prokaryotic genomes. Current Opinion in Microbiology. 11(5) pp. 454–460.

Rocha EP, & Danchin A. 2002. Base composition bias might result from competition for metabolic resources. Trends in Genetics. 18(6) pp. 291–294.

Rocha EPC, Viari A, & Danchin A. 1998. Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. Nucleic Acids Research. 26(12) pp. 2971–2980.

Rolfe R, & Meselson M. 1959. The Relative Homogeneity of Microbial DNA. Proceedings of the National Academy of Sciences of the United States of America. 45(7) pp. 1039–1043.

Rosen GL, Reichenberger ER, & Rosenfeld AM. 2010. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 27(1) pp. 127–129.

Rousu J, Saunders C, Szedmak S, & Shawe-Taylor J. 2006. Kernel-based learning of hierarchical multilabel classification models. Journal of Machine Learning Research. 7 pp. 1601–1626.

Rudi K. 2009. Environmental shaping of ribosomal RNA nucleotide composition. Microbial Ecology. 57(3) pp. 469–477.

Rudner R, Karkas JD, & Chargaff E. 1968. Separation of B Subtilis DNA into Complementary Strands .3. Direct Analysis. Proceedings of the National Academy of Sciences of the United States of America. 60(3) pp. 921–&.

Saeed I, & Halgamuge SK. 2009. The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. BMC Genomics. 10 Suppl 3 pp. S10.

Saeed I, Tang S-L, & Halgamuge SK. 2011. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. Nucleic Acids Research.

Sandberg R, Branden CI, Ernberg I, & Coster J. 2003. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene. 311 pp. 35–42.

Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, & Coster J. 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Research. 11(8) pp. 1404–1409.

Sanger F, & Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology. 94(3) pp. 441–448.

Sanger F, Nicklen S, & Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 74(12) pp. 5463–5467.

Sarawagi S, & Gupta R. 2008. Accurate max-margin training for structured output spaces. Proceedings of the 25th international conference on Machine learning.

Schildkraut CL, Mandel M, Levisohn S, Smith-Sonneborn JE, & Marmur J. 1962. Deoxyribonucleic Acid Base Composition and Taxonomy of some Protozoa. Nature. 196(4856) pp. 795–796.

Schloss PD et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology. 75(23) pp. 7537–7541.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, & Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12(6) pp. R60.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, & Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Meth. advance on.

Sharma VK, Kumar N, Prakash T, & Taylor TD. 2012. Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. PLoS One. 7(4) pp. e34030.

Sharp PM, & Li WH. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15(3) pp. 1281–1295.

Sims GE, Jun SR, Wu GA, & Kim SH. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proceedings of the National Academy of Sciences of the United States of America. 106(8) pp. 2677–2682.

Singer CE, & Ames BN. 1970. Sunlight ultraviolet and bacterial DNA base ratios. Science. 170(3960) pp. 822–825.

Smith TF, & Waterman MS. 1981. Identification of common molecular subsequences. Journal of Molecular Biology. 147(1) pp. 195–197.

Sneath PHA. 1989. Predictivity in Taxonomy and the Probability of a Tree. Plant Systematics and Evolution. 167(1-2) pp. 43–57.

Snel B, Huynen MA, & Dutilh BE. 2005. Genome trees and the nature of genome evolution. Annual Review of Microbiology. 59 pp. 191–209.

Sokal R, & Rohlf J. 1962. The Comparison of Dendrograms by Objective Methods. Taxon. 11(2).

Stanier RY, & Van Niel CB. 1962. The concept of a bacterium. Archiv für Mikrobiologie. 42(1) pp. 17–35.

Stanier RY, & Van Niel CB. 1941. The Main Outlines of Bacterial Classification. Journal of bacteriology. 42(4) pp. 437–66.

Steiger JH. 1980. Tests for comparing elements of a correlation matrix. Psychological Bulletin. 87(2) pp. 245–251.

Stoeckle MY, & Hebert PD. 2008. Barcode of life. Scientific American. 299(4) pp. 82–86,88.

Sueoka N. 1961a. Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harbor Symposia on Quantitative Biology. 26 pp. 35–43.

Sueoka N. 1961b. Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein. Proceedings of the National Academy of Sciences of the United States of America. 47(7) pp. 1141–&.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proceedings of the National Academy of Sciences of the United States of America. 48 pp. 582–592.

Supek F, Skunca N, Repar J, Vlahovicek K, & Smuc T. 2010. Translational Selection Is Ubiquitous in Prokaryotes. Plos Genetics. 6(6).

Suzuki H, Sota M, Brown CJ, & Top EM. 2008. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Research. 36(22) pp. e147.

Suzuki H, Yano H, Brown CJ, & Top EM. 2010. Predicting plasmid promiscuity based on genomic signature. Journal of Bacteriology. 192(22) pp. 6045–6055.

Swartz MN, Kornberg A, & Trautner TA. 1962. Enzymatic Synthesis of Deoxyribonucleic Acid .11. Further Studies on Nearest Neighbor Base Sequences in Deoxyribonucleic Acids. Journal of Biological Chemistry. 237(6) pp. 1961–&.

Takahashi M, Kryukov K, & Saitou N. 2009. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics. 93(6) pp. 525–533.

Teeling H, Meyerdierks A, Bauer M, Amann R, & Glockner FO. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. 6(9) pp. 938–947.

Thomas T, Gilbert J, & Meyer F. 2012. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2(1) pp. 3.

Tringe SG et al. 2005. Comparative metagenomics of microbial communities. Science. 308(5721) pp. 554–557.

Tsochantaridis I, Hofmann T, Joachims T, & Altun Y. 2004. Support vector machine learning for interdependent and structured output spaces. In: ACM: Banff, Alberta, Canada p. 104.

Tsochantaridis I, Joachims T, Hofmann T, & Altun Y. 2005. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research. 6 pp. 1453–1484.

Turnbaugh PJ et al. 2010. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. Proceedings of the National Academy of Sciences of the United States of America. 107(16) pp. 7503–7508.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, & Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 444(7122) pp. 1027–1031.

Tyson GW et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 428(6978) pp. 37–43.

Valouev A et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Research. 18(7) pp. 1051–1063.

Vapnik VN. 1995. *The Nature of Statistical Learning Theory*. Springer.

Vapnik VN. 1998. The Support Vector method of function estimation. Nonlinear Modeling. pp. 55–85.

Venter JC et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 304(5667) pp. 66–74.

Vinga S, & Almeida J. 2003. Alignment-free sequence comparison-a review. Bioinformatics. 19(4) pp. 513–523.

Wang HC, Badger J, Kearney P, & Li M. 2001. Analysis of codon usage patterns of bacterial genomes using the self-organizing map. Molecular Biology and Evolution. 18(5) pp. 792–800.

Wang HC, Susko E, & Roger AJ. 2006. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. Biochemical and Biophysical Research Communications. 342(3) pp. 681–684.

Wang Y, Hill K, Singh S, & Kari L. 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene. 346 pp. 173–185.

Ward BB. 2002. How many species of prokaryotes are there? Proceedings of the National Academy of Sciences of the United States of America. 99(16) pp. 10234–10236.

Warnecke F et al. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature. 450(7169) pp. 560–565.

Watkins C. 2000. Dynamic Alignment Kernels. In: Advances in Large Margin Classifiers. Smola, A & Bartlett, P, editors. MIT Press: Cambridge, MA, USA pp. 39–50.

Watson JD, & Crick FHC. 1953. Molecular Structure of Nucleic Acids - a Structure for Deoxyribose Nucleic Acid. Nature. 171(4356) pp. 737–738.

Webb CO, Ackerly DD, McPeek MA, & Donoghue MJ. 2002. Phylogenies and community ecology. Annual Review of Ecology and Systematics. 33 pp. 475–505.

Weston J, & Watkins C. 1999. Support vector machines for multiclass pattern recognition. In: Proceedings of the Seventh European Symposium On Artificial Neural Networks.

Willenbrock H, Friis C, Juncker AS, & Ussery DW. 2006. An environmental signature for 323 microbial genomes based on codon adaptation indices. Genome Biol. 7(12) pp. R114.

Woese CR, & Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America. 74(11) pp. 5088–5090.

Wu D et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature. 462(7276) pp. 1056–1060.

Wu M, & Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9(10) pp. R151.

Wu Z, Markert K, & Sharoff S. 2010. Fine-grained genre classification using structural learning algorithms. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.

Wu Z, & Palmer M. 1994. Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics: Las Cruces, New Mexico pp. 133–138.

Xu Z, & Hao B. 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Research. 37 pp. W174–W178.

Zeldovich KB, Berezovsky IN, & Shakhnovich EI. 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 3(1) pp. e5.

LIST OF OWN PUBLICATIONS

Book chapters

Alice Carolyn McHardy and Kaustubh Patil, Methods for the phylogenetic binning of metagenome sequence samples, In: F.J. de Bruijn (Editor) Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches, John Wiley & Sons Inc., 2011.

Journal articles

Patil K. R., McHardy A.C., Alignment-free genome tree inference by learning group-specific distance metrics, *in review*.

Patil K. R., Roune L. and McHardy A. C., The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS ONE*, 7(6), 2012.

Patil K. R., Haider P., Pope P. B., Turnbaugh P. J., Morrison M., Scheffer T. and McHardy A. C., Taxonomic metagenome sequence assignment with structured output models , *Nature Methods*, 8(3), 2011, pp. 191--192.