CHARACTERIZING THE DISTINGUISHABILITY OF MICROBIAL GENOMES

by

Scott Cameron Perry

Submitted in partial fulfillment of the requirements for the degree of
Masters of Science

at

Dalhousie University
Halifax, Nova Scotia
April 2010

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Characterizing the Distinguishability of Microbial Genomes" by Scott Cameron Perry in partial fulfillment of the requirements for the degree of Master of Science.

Dated: April 21, 2010

Supervisor: _____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: April 21 2010

AUTHOR: Scott Cameron Perry

TITLE: Characterizing the Distinguishability of Microbial Genomes

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: MSc CONVOCATION: October YEAR: 2010

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

_____

Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

The field of metagenomics has shown great promise in the ability to recover microbial DNA from communities whose members resist traditional cultivation techniques, although in most instances the recovered material comprises short anonymous genomic fragments rather than complete genome sequences. In order to effectively assess the microbial diversity and ecology represented in such samples, accurate methods for DNA classification capable of assigning metagenomic fragments into their most likely taxonomic unit are required. Existing DNA classification methods have shown high levels of accuracy in attempting to classify sequences derived from low-complexity communities, however genome distinguishability generally deteriorates for complex communities or those containing closely related organisms. The goal of this thesis was to identify factors both intrinsic or external to the genome that may lead to the improvement of existing DNA classification methods and to probe the fundamental limitations of composition-based genome distinguishability.

To assess the suite of factors affecting the distinguishability of genomes, support vector machine classifiers were trained to discriminate between pairs of microbial genomes using the relative frequencies of oligonucleotide patterns calculated from orthologous genes or short genomic fragments, and the resulting classification accuracy scores used as the measure of genomic distinguishability. Models were generated in order to relate distinguishability to several measures of genomic and taxonomic similarity, and interesting outlier genome pairs were identified by large residuals to the fitted models. Examination of the outlier pairs identified numerous factors that influence genome distinguishability, including genome reduction, extreme G+C composition, lateral gene transfer, and habitat-induced genome convergence. Fragments containing multiple protein-coding and non-coding sequences showed an increased tendency for misclassification, except in cases where the genomes were very closely related. Analysis of the biological function annotations associated with each fragment demonstrated that certain functional role categories showed increased or decreased tendency for misclassification. The use of pre-processing steps including DNA recoding, unsupervised clustering, 'symmetrization' of oligonucleotide frequencies, and correction for G+C content did not improve distinguishability.

Existing composition-based DNA classifiers will benefit from the results reported in this thesis. Sequence-segmentation approaches will improve genome distinguishability by decreasing fragment heterogeneity, while factors such as habitat, lifestyle, extreme G+C composition, genome reduction, and biological role annotations may be used to express confidence in the classification of individual fragments. Although genome distinguishability tends to be proportional to genomic and taxonomic relatedness, these trends can be violated for closely related genome pairs that have undergone rapid compositional divergence, or unrelated genome pairs that have converged in composition due to similar habitats or unusual selective pressures. Additionally, there are fundamental limits to the resolution of composition-based classifiers when applied to genomic fragments typical of current metagenomic studies.

List of Abbreviations and Symbols Used

ATV	Average tetranucleotide vectors
BLAST	Basic local alignment search tool
CA	Classification accuracy
EBPR	Enhanced biological phosphorus removal
FAMeS	Fidelity of analysis of metagenomic samples
GOS	Global Ocean Sampling
HMM	Hidden Markov model
LCTR	Lowest common taxonomic rank
LGT	Lateral gene (genetic) transfer
NBC	Naïve Bayes classifier
nBLASTP	Normalized BLASTP
NCBI	National Center for Biotechnology Information
NCC	Normalized cut clustering
PCA	Principal components analysis
PCR	Polymerase chain reaction
Pfam	Database of protein families
RAM	Random access memory
RBF	Radial basis function
RBH	Reciprocal best hit
rDNA	Ribosomal DNA
SOM	Self-organizing Map
SVM	Support Vector Machine
TACOA	Taxonomic composition analysis classifier
TFP	Tetranucleotide frequency profiles

C	Cost parameter used when training SVMs
γ	Gamma parameter used when training SVMs
S_n	Sensitivity
S_p	Specificity

Acknowledgements

I would like to express my gratitude to the following people who have helped me to see this thesis through to the end.

To Christian and my fellow students in the Beiko and Blouin labs, past and present - thank you for your lively discussions and moral support along the way. You helped to keep me on track and to ensure that I maintained some degree of balance in my life. It has been a pleasure working with all of you over the past 2.5 years.

To my friends and family - thank you for your unrelenting encouragement and understanding throughout this latest of my endeavors. Time may sometimes keep us apart, but it means a lot to know that I can always count on your support.

To Hannah – I have been redefining 'busy' since the day we met, and you have been nothing both supportive and caring along the way. Thank you for being there, for making me eat and sleep, and for showing genuine interest in my work. You did everything in your power to make the past year easier on me, and I will never be able to thank you enough.

I would like to send a special thank-you to to Andrew Wong for numerous stimulating discussions regarding SVMs and DNA classification during the early days of my program. His input helped to shape the basic idea that inevitably became the theme of my thesis.

To Rob Beiko – This thesis would not have been possible without your support and guidance. Thank you for providing me with the opportunity to attend grad school, for giving me ample flexibility throughout my program, and for mentoring me along the way. I could not have asked for a better supervisor.

Chapter 1 – Introduction

Definition: DNA Classification

In the context of this thesis, DNA classification refers to the attribution of an anonymous DNA sequence to its originating genome or a specific taxonomic unit. Classification may be based upon compositional characteristics of the query sequence, including such features as G+C content or differences in the relative frequencies of short oligonucleotides. Alternatively, in cases where the query sequences contain genes or gene fragments, classification of such fragments may be guided using homology-based approaches that compare each of the anonymous sequences against databases containing genes of known origin.

Motivation for Accurate DNA Classification Methods

In recent years, the need for accurate methods for DNA classification has become increasingly evident, driven largely by the appearance of high throughput DNA sequencing platforms such as the Illumina Genome Analyzer, (http://www.illumina.com) Roche 454 Genome Sequencer (http://www.454.com), and ABI SOLiD sequencer (http://www.appliedbiosystems.com). From a typical DNA sample, these systems generate millions of short (36-600 bp) reads that must subsequently be reassembled in order to reconstruct the source genome or genomes. Traditionally, microbial studies have focused on organismal genomics, involving the isolation of a prokaryotic organism of interest followed by clonal cultivation and Sanger sequencing. Although this traditional approach has proven to be successful for a variety of organisms such as *Haemophilus influenzae* [1], *Escherichia coli* K-12 [2], *Mycoplasma genitalium* [3], and *Bacillus subtilis* [4], in reality very few microbes prove amenable to lab cultivation. This phenomenon was originally described by Staley et al. as the 'great plate count anomaly', when it was observed that plate counts of bacterial cells in culture were often orders of magnitude smaller than the corresponding cell counts for the original samples [5]. More recently, several estimates of microbial susceptibility to lab cultivation suggest that only a minute fraction (0.001% - 3%) of the total microorganisms in existence may be cultured

using existing cultivation methods [6-8]. Advances in lab cultivation techniques and growth media will likely increase the range of organisms for which clonal cultivation is an effective option; however, the complex metabolic and organismal interdependencies that exist within microbial communities may forever limit the applicability of these techniques to the 'unculturable majority'. Even with an increasing arsenal of improved cultivation practices, the study of an organism in isolation greatly reduces our ability to understand the organism's unique role within its microbial community, and sheds little light on the complex biochemical pathways that may in fact span multiple organisms in a given environment [9-11].

With such a large proportion of the microbial diversity and ecology out of reach of traditional cultivation and sequencing methods, there has been a shift toward attempting to study entire communities of microbes in their natural environments, thus removing the requirements for isolation and cultivation of a particular organism of interest. While DNA sequencing was once a time consuming and often cost-prohibitive process, recent advances such as automated Sanger sequencing and massively parallel sequencing by synthesis techniques (Roche 454, Illumina Genome Analyzer) have served to simultaneously increase sequencing throughput and greatly decrease cost. Additionally, the improved sequencing methods require far less input DNA, with 3rd generation sequencing technologies promising to bring about the ability to sequence single DNA molecules while forgoing the current reliance on PCR-based amplification (http://www.pacificbiosciences.com, [12]). The net effect is that DNA sequencing has become more accessible and more widely applicable, and it is now possible to perform shotgun sequencing on DNA extracted from communities of microbes in environmental or clinical samples. This application of shotgun sequencing to entire communities of microbes has led to a new field known as metagenomics, or community genomics.

Metagenomics

Metagenomics, although still a relatively new field, has already proven to be a successful method for studying unculturable organisms from a variety of environments. Two high-profile studies, the Sargasso Sea [13] and Global Ocean Sampling (GOS)

expeditions [14-16], have successfully applied high-throughput metagenomics in order to interrogate the microbial and viral populations from seawater sampled at regular intervals around the globe. In the latter GOS study, analysis of the metagenomic data sets revealed the presence of ~1800 bacterial species spread across 41 open-ocean and coastal sample sites. Also retrieved from the GOS samples were ~6 million new genes that had never before been identified. Another large metagenomics initiative, the Human Microbiome Project [17], involves the application of metagenomics to microenvironments within the human body, and has great implications for helping to identify the role of microbes in both maintaining and degrading human health. Metagenomics has also shown promise in the elucidation of the biochemical pathways involved in important industrial processes. Enhanced biological phosphorus removal (EBPR), a common wastewater treatment method used worldwide in order to decrease the impact of eutrophication, was for many years known only as a 'black box' [18]. The microbes present in the EBPR medium resisted cultivation using traditional techniques, and although specific reaction conditions for the EBPR process were widely understood, very little was actually known about the organism(s) responsible for the process of phosphorus accumulation [19]. Through the use of metagenomics, the genomic complement of the predominant EBPR microorganisms was almost entirely reconstructed, and analysis of the genes found in the various microbes confirmed earlier studies which suggested that Candidatus *Accumulibacter phosphatis* is responsible for phosphorus removal [20-22]. Ongoing metagenomic analyses aim to characterize the relationship between the various strains of *A. phosphatis* and the specific ecologies of the environments from which the genomes were reconstructed. (Slater et al, submitted 2010)

Several variables that are important in determining the success of the metagenomics approach to the interrogation a given microbial community are directly related to the complexity of the community, namely: the total number of species present, the relative abundance of each species, and the phylogenetic and/or compositional relatedness of each of the members. For simple communities containing a small number of well-represented organisms, metagenomics has already shown great promise in reconstructing the component genomes. For instance, in a study of a modest microbial

community found in an acid mine drainage biofilm, near-complete genomes were retrieved for the two dominant members, *Leptospirillum* group II and *Ferroplasma* type II, by simply binning the sequence reads based on similarities in G+C content and sequencing coverage [23]. Additionally, partial genomes were also reconstructed for three less abundant members of the biofilm using the same straightforward approach. For more complex communities, however, much more powerful and as-of-yet unavailable methods of DNA classification will be required if the individual genomes are to be discerned using metagenomics. In the analysis of a large soil metagenome, for example, less than 1% of all of the sequencing reads could accurately be attributed to genomic contigs, limiting downstream analyses to those based solely on the set of genes recovered from the samples rather than complete or nearly complete genomes [24]. A further complication arises from the fact that existing methods for DNA assembly and binning were designed to reconstruct single genomes sequenced from clonal samples, and are insufficiently robust to handle the presence of mixtures of sequences derived from closely related species or those that are highly similar in composition. As such, DNA fragments from these organisms may inadvertently be assembled into chimeric contigs, greatly reducing the utility of the metagenomics approach in examining the interaction and cooperation of the individual members of the given microbial community [25]. Additionally, less-abundant members of a community may fail to achieve adequate representation in the resulting sequencing data, leading only to partial recovery of the associated genomes. [26] In these cases, the attribution of incomplete genomes to the most likely genus or family may still provide insight into the structure of the underlying microbial community, even if the recovery of complete genomes is not achievable using current sequencing techniques [27].

Although the acid mine drainage community has served to validate the use of metagenomics for the study of very simple communities composed of organisms with significant differences in G+C content, complex communities and those containing poorly represented species still remain largely out of reach. Clearly, the development of methods for the accurate attribution of DNA fragments from microbial communities to their true originating genomes or the most likely taxonomic clade should be a key area of

focus if metagenomics is to open the door to the vast genetic diversity present in unculturable microoganisms.

The gold standard for studying the phylogenetic relatedness of a set of organisms has long been based upon the analysis of similarities in 16S rDNA or, less frequently, other highly conserved marker genes such as *recA* [28-31]. Although this marker gene approach might be useful in attempting to classify nearly complete genomic contigs retrieved from low-complexity metagenomes, in practice few marker genes tend to be recovered from complex environmental samples. For example, only 4,125 complete or partial 16S rDNA sequences were recovered from the 7.7 million sequence reads in the Global Ocean Sampling expedition [16]. In some instances, the phylogenetic composition of a microbial community may be interrogated through targeted sequencing of specific marker genes, however these studies generally offer little information as to the distribution of functional roles within the community [32]. Additionally, viruses provide yet another relatively unharnessed avenue for the discovery of novel genetic diversity, however, viral genomes tend to be extremely compact and do not contain the equivalent of bacterial 16S rRNA genes, thus limiting the utility of the marker gene approach. Alternatives to the gold standard include classification methods that rely upon the homology of environmental DNA fragments to sequences present in databases of known genomes [33-35], as well as techniques that attempt to classify DNA fragments based on compositional characteristics that may be specific to a particular genome or taxonomic clade [36-38].

Genome Signature

It has been well established that between-genome compositional variation for a pair of microbes tends to be significantly higher than within-genome compositional variation, especially for pairs of genomes that are separated by a large phylogenetic distance [39-42]. G+C composition is perhaps the simplest measure of such compositional variation, and as indicated in the preceding section, can be sufficient to discriminate between genomic fragments from multiple organisms in at least some low-complexity microbial communities [23]. This pattern of within-genome composition has

often been referred to as the genome signature. Factors implicated in the establishment of an organism's specific genome signature include biases induced by DNA replication and repair mechanisms [39; 43], codon usage [44], avoidance of restriction endonuclease cleavage sites [43; 45], growth environment [46], and DNA base stacking conformation [43]. Collectively, these forces serve to shape the composition of an organism's genome, providing a signal which may be harnessed in order to assign metagenomic or otherwise anonymous DNA fragments to the correct source genome or taxonomic group.

Early studies into genome signature dealt largely with G+C composition as well as various measures of codon usage bias such as the codon adaptation index [47]. Although these measures of genome signature are sufficient to discriminate between pairs of genomes in some instances, both measures are susceptible to crowding of the feature space, and have been shown to carry little phylogenetic signal [40; 48; 49]. Karlin et al. first reported an improved measure of genome signature based on the over and underrepresentation of dinucleotides present in a genome's DNA sequence, and showed that this dinucleotide relative abundance was a more effective means of discriminating between microbial genomes [39; 43; 50]. Building upon this work, several other authors demonstrated that the relative frequencies of longer oligonucleotides also captured species-specific compositional features as well as phylogenetic signal [51]. In particular, the frequencies of tetranucleotides observed in genomic DNA have successfully been applied to the unsupervised clustering of DNA fragments into compositional bins [40; 45] as well as the attribution of DNA fragments into taxonomic clades of varying levels [36-38; 52]. In situations where marker gene approaches are not an option, techniques based upon genome signature may provide an effective means for determining the origin of anonymous DNA fragments.

It is now widely understood that genes may also be passed laterally between organisms belonging to different species, sometimes over great phylogenetic distances. This phenomenon, known as lateral genetic transfer (LGT) or horizontal gene transfer, allows for the rapid acquisition of genes that might encode features such as antibacterial resistance or other important biosynthetic pathways. Lateral genetic transfer between

compositionally divergent microbes is likely to increase the within-genome compositional variation of the acceptor organism, and therefore has the potential to obscure the genome signature in the vicinity of the introgressed sequence. Although most prevalent in bacteria and archaea, evidence also suggests that eukaryotes may also be susceptible to lateral gene transfer, albeit at a much slower rate than observed in the prokaryotic world [53; 54].

As a species evolves over many generations, any sequence acquired via LGT will gradually change in composition and converge toward the genome signature of the host through the process of amelioration [55]. Given sufficient evolutionary time, the compositional characteristics of laterally transferred genes will eventually become indistinguishable from those inherited vertically, as DNA replication biases and other factors that influence an organism's genome signature slowly bring the composition of the foreign genes in line with the acceptor organism's genome. This poses an immense challenge for both composition and homology-based DNA classification methods, as genes recently acquired via LGT may easily be mis-attributed to the donor organism. In fact, numerous surrogate methods have been employed in order to attempt to identify genes implicated in LGT by searching for regions of a given genome with compositional signatures that differ from the predominant genome signature [52; 56-58].

Review of Existing Methods for DNA Classification

Existing methods for DNA classification can be grouped into one of three classification paradigms: *supervised methods* that necessitate some form of labelled reference data in order to train a machine-learning algorithm or to act as a comparator data set, *unsupervised methods* that perform classification based entirely on characteristics intrinsic to the test data, and *semi-supervised methods* that share aspects of both the supervised and unsupervised paradigms. Depending on the manner in which a classifier represents and compares DNA fragments, the various DNA classification methods can be further grouped into two additional categories. *Sequence-composition* methods rely on the innate compositional characteristics of a given DNA sequence such as G+C content, codon usage biases in the case of coding regions, and oligonucleotide

frequency profiles (essentially DNA word frequencies) in order to facilitate classification. Such methods may include supervised classifiers that depend upon a reference corpus of known genomes in order to build models capable of recognizing genome signatures specific to certain phylogenetic clades. Alternatively, in the absence of a database of reference genomes, several unsupervised sequence-composition classifiers are able to cluster anonymous DNA fragments into compositional bins in a phylogenetically naive fashion based solely on the properties of the DNA fragments themselves. In contrast, *sequence-similarity* or *sequence-homology* based classifiers must compare query sequences against databases of known genes and/or genomes and subsequently utilize the various similarity measures in order to classify sequences into specific clades. The dependence of homology-based methods on reference sequences necessitates that such methods fall into the supervised category of classifiers.

A classification method's reliance on databases of known sequences may have both positive and negative implications. Classifiers that rely entirely on comparisons against public reference databases may succeed in classifying only those anonymous fragments that have close relatives in the training data set, which at present contains but a small fraction of the microbial diversity present in nature [36; 59]. In contrast, classifiers that bin sequences without the aid of external databases are oblivious to existing phylogenetic clades, and as a consequence require post-classification manual intervention in order to assign the resulting clusters of fragments into the existing phylogenetic hierarchy [40; 59]. Clearly, the choice as to which type of classifier is most appropriate will depend largely on the nature of a given experimental data set, and the overall relatedness of the query sequences to the sequences present in the various reference databases.

Several model examples of existing techniques for DNA classification and binning will be examined in the following sections. The complete list of methods is summarized in Table 1.1.

Supervised Training of Classifiers

Chi-squared Approach

The Chi-squared classification method (referred to as the *k-mer* method by the authors) is a relatively simple sequence-composition classifier developed by the US Department of Energy's Joint Genome Institute (DOE-JGI) and described in the Fidelity of Analysis of Metagenomic Samples (FAMeS) paper by Mavromatis et al. [25]. The FAMeS manuscript presents three simulated metagenomic datasets, each designed to represent a different level of community complexity in terms of the number and relative abundance of unique microbial populations present in the sample. A low complexity metagenome consists of one well-represented organism surrounded by a small number of low-abundance organisms, such as the enhanced biological phosphorus removal (EBPR) bioreactor metagenome [20]. In contrast, a high complexity metagenome lacks any single dominant population, and instead consists of numerous organisms that are poorly represented in the sample, such as metagenomic samples obtained from soil [60]. The goal of the FAMeS manuscript is to compare the classification performance of the Chi-squared method versus two additional classifiers (PhyloPythia [38] and BLAST distribution [25]) when applied to three simulated metagenomic datasets of increasing community complexity (simLC: low complexity, simMC: medium complexity, and simHC: high complexity). Both PhyloPythia and BLAST distr will be discussed at length in subsequent sections.

In order to evaluate the Chi-squared classifier, all of the reference genomes included within version 1.3 of the Integrated Microbial Genomes system [61], minus the dominant members of the simLC, simMC, and simHC metagenomes, were used to construct the training set. Each reference genome was first partitioned into fragments 8000bp in length, and each 8000bp subsequence was represented by the relative frequencies of all possible overlapping 7-mers and 8-mers present on either strand in the fragment. Similarly, the test set consisted of all fragments >= 8000bp in length from three distinct assemblies of the simLC, simMC, and simHC simulated metagenomic datasets. All test fragments were likewise represented using their corresponding 7-mer and 8-mer

frequency profiles.

Classification was facilitated by comparing the oligonucleotide frequency profiles of each of the test fragments against the entire set of reference frequency profiles using either of the following two patterns: “NNNNNNN” or “NNxNNxNN” where N represents any nucleotide and x is ignored during the pattern-matching step. Comparisons between fragments were performed using the oligonucleotide frequency profiles associated with both strands of DNA. Each fragment was then assigned to the taxonomic family of the best matching reference genome, according to a Chi-squared comparison.

Overall, the performance of the Chi-squared method was quite poor in relation to the other methods examined in the FAMeS study, with average specificity <= 11% and average sensitivity <= 24% across the three assemblies and three simulated data sets. Although this method failed to classify fragments into the correct taxonomic family in the vast majority of cases, the authors noted that the method was consistently capable of binning fragments into the correct taxonomic order, despite the poor performance at the family level. Refinement of the method, perhaps using shorter length oligonucleotides in calculating the k-mer frequency profiles, may lead to improved performance at more specific taxonomic ranks.

Table 1.1: Summary of Existing Methods for DNA Classification and Binning

Coloured shading is used to highlight the machine learning and classification strategy used by each method. **blue**: supervised composition-based classifiers, **yellow**: unsupervised composition-based classifiers, **orange**: semi-supervised composition-based classifiers, **purple**: supervised homology-based classifiers.

Classifier	Category of machine learning	Classification strategy	Methodology	Appropriate fragment length	References
PhyloPythia	Supervised	Composition	Hierarchy of multiclass SVMs trained using tetranucleotide frequency profiles at various taxonomic ranks	1kbp - 50kbp	[38]
Naïve Bayesian	Supervised	Composition	Probabilistic classification of n-mer frequency profiles using a Bayesian classifier	400bp – 1kbp, 25bp - 500bp	[52; 62]
TACOA	Supervised	Composition	Oligonucleotide frequencies clustered using k-nearest neighbor algorithm combined with a Gaussian kernel	800bp - 50kbp	[37]
Chi-squared (FAMeS)	Supervised	Composition	Comparison of 7-mer or degenerate 8-mer oligonucleotide frequency profiles to those of known genomes using Chi-squared measure	8kbp+	[25]
TETRA	Unsupervised	Composition	Fragments binned based on pairwise Z-score correlations of tetranucleotide frequencies	40kbp+	[40; 63]
SOM	Unsupervised	Composition	Clustering of tetranucleotide frequencies into anonymous phylotypes using a SOM	5kbp+	[64]
S-GSOM	Semi-supervised	Composition	GSOM with post processing to cluster sequences based on seeds (16S flanking sequences)	8kbp - 13kbp	[36]
CompostBin	Semi-supervised	Composition	Weighted-PCA of hexanucleotide frequencies combined with a recursive normalized-cut clustering algorithm. Clustering is augmented using phylogenetic markers.	~1kbp+	[65]
BLAST distr (FAMeS)	Supervised	Homology	Genes predicted using fgenesb and normalized-BLASTP best hits are subsequently used to assign fragments into the most likely taxonomic class	8kbp+	[66]
CARMA	Supervised	Homology	Conserved PFAM domains and families found in query sequences are used to classify sequences into specific taxonomic ranks	80bp - 400bp	[33]

Naïve Bayes Classifier

The naïve Bayes classifier (NBC) is an application of Bayesian statistics to classification, under the 'naïve' assumption that the specific set of features that define any given class in a multi-class problem are completely independent of one another. Although the assumption of feature independence is often violated in practice, the naïve Bayes classifier has proven to be robust to such violations for a number of applications, including text classification [67], the prediction of protein function [68], and spam filtering [69].

Two separate studies have used oligonucleotide frequency profiles in conjunction with the probabilistic naïve Bayes classifier in order to classify DNA fragments from sets of bacterial and archaeal genomes [52; 62]. In both studies, genomes were first partitioned into sets of fragments of assorted lengths, and k-mer frequency profiles were subsequently calculated for each fragment length using various values of k. The k-mer frequency profiles were then classified by the naïve Bayes classifier in a cross-validated fashion, and the performance of the classifier was measured in terms of its global classification accuracy for each of the possible fragment length and k-mer length combinations.

Sandberg et al [52] examined NBC performance using a set of 28 bacterial and archaeal genomes, with fragment lengths ranging from 35 nt – 1000 nt, and k-mer frequency profiles calculated over all possible oligonucleotide lengths up to a maximum of 9 nt. In each trial, the training set consisted of oligonucleotide frequency profiles generated from 100 randomly selected fragments from each genome. Classifier performance on the training set was evaluated for each combination of fragment length and k-mer pattern length using 10 fold leave-one-out cross-validation. The authors noted that classification accuracy increased with both increasing fragment length and increasing k-mer length, and the NBC achieved a maximum classification accuracy of nearly 90% using 9-mer frequency profiles calculated from 1000 nt genomic fragments. Interestingly, even very short fragments could often be classified by the NBC, with 35 nt fragments leading to a classification accuracy of 36% (compared with a baseline accuracy of 3.57%

for random predictions), and 60 nt fragments resulting in a classification accuracy of 46%, both using 9-mer frequency profiles. For the second longest fragment length, 400 nt, the NBC achieved a maximum classification accuracy of 85%, once again using 9-mer frequency profiles.

In a more comprehensive study, Rosen et al applied the NBC approach in order to classify fragments from 635 bacterial and archaeal genomes using fragment lengths of 25 nt, 100 nt, and 500 nt, and k-mer lengths ranging from 3 nt – 15 nt [62]. For each fragment length and k-mer length combination, the training set was constructed by partitioning each genome into substrings of the appropriate fragment length, and calculating the corresponding k-mer frequencies for all fragments. As with the Sandberg study, the authors noted that classification accuracy tended to increase with increasing fragment and k-mer lengths. Interestingly, it was noted that for very short fragments, optimum k-mer length was inversely proportional to the fragment length. For instance, for 100 nt and 500 nt fragments, classification accuracy appeared to plateau using 12-mer frequency profiles, whereas the highest classification accuracy for 25 nt fragments was achieved using 15-mer frequency profiles. As k-mer length increases, the corresponding feature vectors become increasingly sparse, such that most features will have no representation, and those features with a frequency >1 will likely be specific to a given species or genus. For this reason, the observation that longer k-mers resulted in increased classification accuracy for shorter fragments may simply be artefacts whereby the classifier is recognizing the primary nucleotide sequence of each fragment rather than the compositional characteristics of the fragment.

The authors reported maximum species-level classification accuracies of 97.3% for 500 nt fragments, 95.3% for 100 nt fragments, and 90.2% for 25 nt fragments using 5-fold cross-validation on a reduced subset of genomes. It should be noted, however, that these results must be interpreted in the context of the cross-validation methodology employed in the study. In order to calculate the cross-validated classification accuracies, the authors first selected a subset of 77 strains of bacteria/archaea, representing 9 unique species. Five-fold leave-one-out cross-validation was performed by randomly partitioning

the 77 strains (rather than partitioning the k-mer frequency profiles from the fragments associated with each strain) into 5 cross validation groups. Furthermore, the classification accuracies for the cross-validation trials were reported at the species level rather than the strain level upon which the cross-validation procedure was based, thus relaxing the complexity of the classification problem from 77 classes to 9.

Performance of the NBC at the strain level was calculated by training the classifier using the complete set of k-mer frequency profiles from all 635 genomes, and then testing the classifier using the n-mer frequency profiles associated with 100 randomly selected fragments from each genome. Using this methodology, the classifier achieved strain-level classification accuracies of 88.8% for 500 nt fragments, 82.5% for 100 nt fragments, and 75.8% for 25 nt fragments. Since the test fragments were present in both the testing and training sets (no cross-validation was employed in this case), the strain-level performance values may have been inflated due to overfitting of the model.

PhyloPythia

PhyloPythia is a metagenomic classification system designed to bin short DNA fragments into relevant phylogenetic clades based upon pentanucleotide and hexanucleotide frequency profiles using a multi-class support vector machine (SVM). [38] The training set used to measure the performance of PhyloPythia represented 340 completely sequenced bacterial and archaeal genomes. The method was used to classify fragments of various lengths {1kbp, 3kbp, 5kbp, 10kbp, 15kbp, 50kbp} at each of the taxonomic ranks of domain, phylum, class, order, and genus. As with the classifiers previously discussed, PhyloPythia generated models and performed classifications based on the G+C- and length-corrected oligonucleotide frequency profile representations of short genomic fragments. Several preliminary analyses by McHardy et al [38] heavily influenced the overall design of PhyloPythia. Notably, it was shown that different oligonucleotide pattern lengths were most appropriate for different taxonomic ranks, with pentanucleotide patterns performing optimally for the more specific ranks of class, order, and genus, while hexanucleotide patterns resulted in the best classification at the more general ranks of domain and phylum (although there was only marginal improvement in

classification accuracy for k-mers > 4 nucleotides in length). Additionally, it was shown that the classification accuracy of a given query fragment is largely influenced by the difference between the length of the query fragment and the length of the fragments used in the construction of the SVM model. More specifically, it was observed that the classifier performed optimally when the training and query fragments were of comparable length, whereas classifiers trained using fragments longer than the query fragment showed decreased performance in proportion to the training fragment length. Classification accuracy deteriorated rapidly for the cases where the query fragments were longer than the training fragments, although in practice this effect might be mitigated by attempting to classify shorter subsequences of the query fragments.

With the aforementioned parameters in mind, PhyloPythia was designed as a large array of hierarchical SVMs, where a distinct SVM was trained for each fragment length and taxonomic rank combination. Within each SVM, phylogenetic clades containing >= 3 genomes were represented as individual classes, while all clades with fewer than 3 genomes were pooled to generate an 'other unknown' class. For each SVM, the training data were represented by the appropriate pentanucleotide or hexanucleotide frequency profiles, as discussed above. Additionally, for each well-represented phylogenetic clade present in the training set, a one-against-the-rest SVM was trained for each fragment length and taxonomic rank combination, with the genomes from a single phylogenetic clade representing one class, and all genomes from all other phylogenetic clades representing a second class. This latter set of SVMs was used in a post-processing step in order to validate the initial SVM predictions in an attempt to reduce the incidence of false positives.

At each taxonomic rank, classification of a query fragment is achieved by sequentially passing the query fragment through the hierarchy of SVMs in order of decreasing length of the training fragments, until the query fragment is successfully classified into a specific clade or the length of the query fragment is longer than the fragments used to train the next available set of SVMs. When a fragment has been classified into a specific clade, the fragment is subsequently passed through the

appropriate post-processing SVM in order to support or invalidate its assignment to the given clade.

The performance of PhyloPythia in classifying fragments from the training set was first evaluated using a leave-one-out cross-validation strategy, where each genome in turn served as the query genome, while all remaining genomes were used to construct the various models in the SVM hierarchy. This approach was designed to mimic the situation where a metagenome contains a genome that has not yet been observed, and as such is not present in the classifier's training set. Overall, PhyloPythia achieved high accuracy in this evaluation, exhibiting specificities between 79%-96% across all fragment length and taxonomic rank combinations. Sensitivity scores showed a pronounced dependence on query fragment length across all taxonomic ranks, particularly for fragments less than ~5kbp in length. For example, 1kbp fragments achieved the minimum sensitivity of 4.42% at rank genus, while the maximum observed sensitivity was 92.23% for 50kbp fragments, also at rank genus. For fragments >= 5kbp in length, sensitivity never fell below 79.53% across all taxonomic ranks.

In a second evaluation, the classification procedure was repeated for all genomic fragments while omitting the cross-validation procedure, in order to evaluate the performance of PhyloPythia when faced with fragments derived from organisms with genomes that are present in the training set. In this evaluation, it is important to note that despite the inclusion of all genomes in the training set, a proper cross-validation strategy was employed such that there was no overlap between the training and testing sets at the fragment level, i.e, a portion of the fragments from each genome were used as test fragments, while the remaining fragments were used to train the SVM models. As with the previous evaluation, PhyloPythia showed relatively consistent specificities, ranging from 83.66% (10kbp fragments at rank genus) to 99.95% (50kbp fragments at rank domain) across all fragment lengths and taxonomic ranks. Sensitivities once again varied in proportion to query fragment length, ranging from 7.11% (1kbp fragments at rank genus) to 99.8% (50kbp fragments at rank domain). For fragments >= 5kbp in length, sensitivities were comparable across all taxonomic ranks, with no observed sensitivity

below 95.43%. At the rank of genus, there was a dramatic increase in sensitivity between 1kbp fragments (7.11%) and 3kbp fragments (69.16%), indicating that even for the existing set of known genomes it is desirable to have fragments >1kbp in length if reasonable classification accuracy is desired.

The authors compared the performance of PhyloPythia against two other classifiers: 1) a TETRA-like method [40; 63], and 2) a classifier based on the self-organizing map (SOM) [36; 45; 64; 70]. Both TETRA and several derivatives of the SOM method will be examined individually in subsequent sections. Each method was evaluated in terms of its ability to correctly classify fragments associated with the dominant populations present in the Sargasso Sea metagenome [13]. In the classifier comparison, the data set consisted of DNA fragments from metagenomic contigs that contained annotated 16S rRNA genes, such that the various classifier predictions could be directly compared against the presumed phylogenetic identities of fragments associated with each contig. For the purpose of this study, the PhyloPythia models were extended to include 100kb – 162kb of sequence from the four most prevalent bacterial populations present in the Sargasso sample, namely *Prochlorococcus*, unknown *Gammaproteobacteria*, *Shewanella*, and *Burkholderia.* Although the three methods exhibited comparably high specificities, ranging from 94% to 100%, they largely disagreed in terms of the percentage of correctly assigned fragments. At the species level, PhyloPythia successfully assigned 72% of fragments into the correct genomic bin, whereas the TETRA-like method achieved only 39% accuracy. Likewise, in a class-level comparison (the most specific taxonomic rank at which the SOM method was applicable) PhyloPythia correctly assigned 74% of fragments, while the SOM method classified only 20% of fragments correctly in this case. It was noted that PhyloPythia was better suited at classifying shorter fragments, correctly classifying fragments as short as 1.5kbp, whereas the minimum fragment length correctly assigned by TETRA was 12kbp (data not provided for SOM).

TACOA

The Taxonomic Composition Analysis classifier, known as TACOA, demonstrates

that accurate classification of short genomic fragments is possible using even the most simplistic of machine learning algorithms [37]. This classifier leverages the k nearest neighbour (k-NN) algorithm in order to classify DNA fragments based on their underlying oligonucleotide frequency profiles. It is understood that the performance of the traditional k-NN algorithm degrades as the dimension of the feature space increases, an effect known as the 'curse of dimensionality' or Hughes effect. In order to reduce the impact of the curse of dimensionality, Diaz et al chose to augment the k-NN algorithm with a Gaussian kernel density function. This algorithmic modification lessens the impact of the curse of dimensionality by decreasing the weight of the reference oligonucleotide vectors in proportion to their Euclidean distances from the query vector. An added advantage of the Gaussian kernel is that the entire reference set of frequency vectors can be examined during the testing phase, rather than considering only those features that fall within the immediate neighbourhood of the given query vector.

A reference set comprised of 373 completely sequenced bacterial and archaeal genomes was used in the evaluation of TACOA. Each genome was represented by a set of vectors of oligonucleotide frequencies, corrected for both genome length and G+C content, where the oligonucleotide patterns ranged in length from 1-6bp. Performance of the classifier was evaluated using a leave-one-out cross validation strategy, whereby each genome in turn served as the query/test genome, while the oligonucleotide frequency vectors for the remaining 372 genomes formed the training set. For each of the 373 cross validation trials, 3000 non-overlapping genomic fragments were selected at random from the query genome for each of the examined fragment lengths {800bp, 1000bp, 3000bp, 10kbp, 15kbp, 50kbp}, and distinct sets of oligonucleotide frequency vectors were determined using patterns of lengths 1 - 6bp. Classification accuracy was subsequently determined for each fragment using all possible fragment length and oligonucleotide pattern length combinations, across each of the taxonomic ranks superkingdom, phylum, class, order, and genus.

Classification accuracy of the TACOA classifier appeared to be directly influenced by the length of the query fragments, with performance increasing in

proportion to fragment length. For instance, for 800bp fragments, average sensitivity ranged from 5% at taxonomic rank genus to 67% for the rank of superkingdom, and average specificity ranged from ~59% at rank genus to ~75% at rank superkingdom. Comparatively, these values are significantly lower than those of 50kb fragments, which had average sensitivities ranging from 46% to 82%, and average specificities ranging from 77% to 93%, for the same set of taxonomic ranks. Overall, the classifier showed a low rate of misclassification, with a false negative rate of 10% or lower across all fragment lengths and taxonomic ranks considered. The authors also noted that tetranucleotide frequency vectors were most appropriate for the classification of fragments <= 3000bp in length, whereas 10kbp, 15kbp, and 50kbp fragments were best classified using pentanucleotide frequency vectors. Interestingly, it was also demonstrated that the use of oligonucleotide patterns greater than 5bp in length resulted in a decrease in both the average specificity and sensitivity, and an increase in the false negative rate across all fragment lengths and taxonomic ranks.

In a separate analysis, the performance of TACOA was compared directly to that of PhyloPythia [38] using a test set consisting of 63 newly sequenced microbial genomes absent from both the TACOA and PhyloPythia reference sets. In this comparison, sensitivity, specificity, and false negative rates were calculated for the results of trials using 3 separate fragment lengths (800bp, 1kbp, 10kbp) and the same five taxonomic ranks previously considered: superkingdom, phylum, class, order, and genus. For the 3 least specific taxonomic ranks (superkingdom, phylum, class), performance of the two classifiers was comparable. In terms of sensitivity, TACOA marginally outperformed PhyloPythia for fragments of length 800bp and 1kbp (except for 800bp fragments at rank class), with both classifiers achieving sensitivities between 66% - 76% for superkingdom, 15% - 28% for phylum, and 3% - 11% for class. PhyloPythia consistently showed higher sensitivities for 10kbp fragments, and significantly outperformed TACOA for 10kbp fragments at ranks phylum (61% vs. 41%) and class (47% vs. 30%). Both classifiers demonstrated decreasing sensitivity for the more specific taxonomic ranks, with longer fragments resulting in the highest sensitivities. Specificities were comparable for both classifiers across all fragment lengths for the ranks of superkingdom, phylum, and class,

ranging from 65% (PhyloPythia: 800bp fragments, rank superkingdom) to 97% (TACOA: 10kbp fragments, rank superkingdom). False negative rates were likewise comparable for 800bp and 1kbp fragments, although PhyloPythia showed much higher false negative rates for 10kbp fragments at ranks phylum (15% vs. 5.33%) and class (27% vs. 7.4%).

For the more specific taxonomic ranks of order and genus, PhyloPythia failed to correctly classify any fragments across any of the considered fragment lengths. In contrast, TACOA achieved low sensitivities ranging from 3% (800bp fragments at rank genus) to 17% (10kbp fragments at rank order) and specificities ranging from 67% (1kbp fragments at rank genus) to 96% (10kbp fragments at rank order). TACOA had low false negative rates at these two ranks, ranging from 1% to 2.43% across the 3 fragment lengths considered in the study.

BLAST Distribution

The BLAST distribution classifier (BLAST distr) is a simple BLASTP [71; 72] based approach to metagenomic binning, originally presented in the Fidelity of Analysis of Metagenomic Samples (FAMeS) paper by Mavromatis et al [25]. As previously described for the Chi-Squared method, BLAST distr was used as a comparator in the FAMeS study in order to evaluate the performance of various metagenomic binning methods when applied to three simulated metagenomic datasets of increasing complexity: simLC (low complexity), simMC (medium complexity), and simHC (high complexity).

The overall premise of the BLAST distr method is to perform BLASTP searches for all proteins identified in a metagenomic sample, and then attempt to assign each metagenomic fragment to a specific phylogenetic clade based on the distribution of its genes' highest-scoring BLASTP hits. In the FAMeS study, the three metagenomic datasets were first analyzed using fgenesb (http://softberry.com) in order to detect genes located on any of the associated fragments >= 8kbp in length. For each of the predicted genes, the relevant protein products were then used as query sequences in BLASTP searches against 253 completely sequenced bacterial and archaeal genomes, with the

exclusion of the dominant members of the simulated metagenomes. Normalized BLASTP scores were determined for any BLASTP hits with expectation values less than the threshold of 1e-05. Each query fragment was then assigned to the taxonomic class with the highest overall normalized BLASTP score, so long as at least 50% of the genes present on the given fragment had BLASTP hits to the relevant class, and the average normalized BLASTP score per gene was > 0.2.

Although the BLAST distr method was only required to predict each fragment's identity at the general level of the most relevant taxonomic class, the method still performed quite poorly for the simMC data set. For this medium complexity metagenome, BLAST distr achieved a maximum sensitivity of 58% and maximum specificity of 59%, whereas PhyloPythia was able to achieve nearly 100% sensitivity and specificity in some instances. BLAST distr showed improved performance on the low complexity data set, however, achieving 100% specificity and 80% sensitivity in this case. As the BLAST distr is directly influenced by the presence (or lack thereof) of closely related sequences in the BLAST databases, it may be expected that the performance of this method will increase as new organisms are sequenced and the reference databases become more comprehensive.

CARMA

Krause et al devised a novel DNA classification system, CARMA, for classifying very short metagenomic fragments into relevant phylogenetic clades through a combined sequence-homology and phylogenetic approach [33; 73]. This method depends heavily on the identification of known protein domains within metagenomic query sequences in order to facilitate classification, a criterion that often limits the applicability of the method to a small fraction of the total reads in a given dataset. Despite this limitation, the authors demonstrated that CARMA is capable of accurately classifying very short reads in which identifiable PFAM domains are present, providing for the potential characterization of the taxonomic diversity of metagenomic datasets that largely consist of unassembled reads.

CARMA performs classification on a read-by-read basis using a multi-step pipeline that includes homology searches, sequence alignments, and the construction of phylogenetic trees. During an initial data-filtering step, a BLASTX [71; 72] search is performed between the metagenomic reads and the entire PFAM [74; 75] database in order to identify the set of reads that are likely to contain complete or partial protein domains curated within PFAM. The BLASTX search is performed using moderately relaxed settings with the intention of detecting all likely PFAM hits, while filtering out those reads that are unlikely to contain conserved protein domains. This step is necessary in order to reduce the number of reads that are included in the subsequent and much more computationally intensive steps of the pipeline. The reads identified as being likely to contain conserved protein domains are next passed through a validation step, whereby each read is searched using a sensitive hidden Markov model (HMM) specific to the PFAM domain family for which the read was matched during the initial BLASTX search. As opposed to the relaxed BLASTX search, the HMM search is performed using a strict E-value cutoff of 0.01 in order to limit the incidence of false positives in the resulting data set. Next, for each of the PFAM families that match one or more reads during the HMM search, a multiple sequence alignment is generated using all PFAM protein sequences from the family along with the protein sequences coded for by each of the reads that matched the given family. Pairwise distance matrices are then calculated from these multiple sequence alignments, and the distance matrices are then used to construct unrooted phylogenetic trees via the neighbor-joining method from the PHYLIP [76] package. Classification of reads is ultimately based upon the specific clustering of the nodes within the resulting phylogenetic trees. If the node representing a given read is contained within a subtree in which the sister PFAM nodes all belong to the same taxon from the NCBI taxonomy database, the read is assigned to that taxon. In the event that the PFAM nodes in the subtree represent multiple taxa, the read is assigned to an 'unknown taxon' class.

In order to evaluate the performance of CARMA, a synthetic metagenome was first created by simulating short 80-120bp reads from 77 bacterial/archaeal genomes at 2X coverage using the ReadSim package [77]. This simulated metagenome was intended

to represent a moderately complex metagenomic community consisting of 62 genera spread across 10 bacterial/archaeal phlya. CARMA was used to classify the reads from this simulated metagenome, while ensuring that all of the 77 test genomes were excluded from the PFAM database during these trials. Upon being presented with the synthetic metagenome, CARMA identified conserved PFAM domains in approximately 15% of the metagenomic reads. Of this 15% of reads, CARMA exhibited reasonable average sensitivities ranging from 61% at the rank of order to 84% at superkingdom, with corresponding specificities ranging from 90% - 97%. Across all taxonomic ranks, CARMA exhibited a relatively consistent false negative rate of approximately 7%, while the false positive rate for each taxonomic group tended to vary in proportion to the number of sequences representing the given taxon in the PFAM database.

CARMA was also used to estimate the taxonomic composition of a relatively low-complexity metagenome from an agricultural biogas reactor [78]. Although this metagenomic dataset actually consists of approximately 600,000 short reads with an average read length of 230bp, the authors decided to simulate ultra-short reads by considering non-overlapping substrings of the original reads. As such, the 600,000 original reads were used to generate 9 separate sets of ultra-short reads, with lengths of 35bp, 40bp, 50bp, 60bp, 70bp, 100bp, 150bp, 200bp, and 250bp. After processing each of the sets of reads through CARMA, it was very apparent that read length had a large influence on the sensitivity of the underlying homology searches. For example, the number of PFAM domains detected in each set of reads was highly influenced by read length, and ranged from 886 for the set of 35bp reads to 89,979 for the set of 250bp reads. While the sensitivity of the CARMA method tended to decrease for the more specific taxonomic ranks, remarkably the proportion of PFAM-containing reads that could not be classified into a specific taxon did not vary to a considerable degree across all fragment lengths considered. For instance, between 9-11% of PFAM-containing reads could not be classified at the level of superkingdom, 43-52% at the level of order, and 57-73% at the level of species. The method was also remarkably consistent in predicting the relative abundance of taxa for each set of reads across all taxonomic ranks. Even at the species level, the relative abundance of each species as predicted by CARMA was shown

to be relatively consistent between the 35bp and 250bp reads.

Unsupervised Training of Classifiers

TETRA

TETRA was one of the earliest methods developed for comparing anonymous DNA fragments based upon tetranucleotide frequency profiles [40; 63]. Although TETRA lacks the ability to classify fragments into existing phylogenetic clades, the method is capable of determining the pairwise compositional relatedness of a given set of fragments, and as such can be applied to metagenomic data sets in order to bin fragments based on similarities in their compositional characteristics. The goals of the TETRA study were to demonstrate that a tetranucleotide-based binning approach is capable of outperforming methods based on fragment G+C content, and to show that TETRA may be useful in helping to bin large, fosmid-sized (40 kbp) fragments from low complexity metagenomes.

For a given set of DNA fragments of size n, TETRA produces $\binom{n}{2}$ pairwise z-score correlations that may be used to help interpret the relatedness of each of the fragments. The z-scores are calculated by first determining the observed tetranucleotide frequencies along both strands of each DNA fragment, and subsequently calculating the expected tetranucleotide frequencies based on a maximal order Markov model. The sets of observed/expected frequencies for each possible tetranucleotide are then converted to z-scores using an approximation method described by Schbath [79]. Finally, for each fragment pair, Pearson's correlation coefficient is calculated from the associated tetranucleotide z-scores. In the ideal case, intragenomic z-score correlations will be significantly higher than intergenomic z-scores, thus allowing compositionally similar fragments to be binned together despite the fact that the phylogenetic identities of the individual bins are unknown. Similarly, fragments may also be binned based upon significant differences in intragenomic and intergenomic G+C, where intragenomic fragments are expected to show less variation in G+C than their intergenomic counterparts.

In order to evaluate the binning performance of TETRA on an artificial fosmid-based data set and compare the results with a common binning method based on G+C composition, 118 completely sequenced bacterial genomes were first partitioned into a set of 40kbp fosmid-sized fragments, representing 9054 fragments in total. For each pair of fragments in the reference corpus, the tetranucleotide z-score correlation (see above) and the difference in G+C composition were calculated. The results were subsequently summarized at the taxonomic ranks of domain, phylum, class, order, and species.

In nearly all cases, TETRA outperformed the G+C binning method in terms of its ability to bin fragments to the correct genome. For instance, 92.7% of all genome pairs had at most 35% nonassignable fragments using the TETRA method, whereas only 74.3% of genome pairs had an equivalent percentage of nonassignable fragments when the G+C binning method used. For a small number of genome pairs, both methods were completely unable to successfully assign fragments, with TETRA failing to discriminate between fragments for 1.4% of genome pairs, while the G+C method failed for 6.7% of all genome pairs. Overall, the results suggest that crowding of the G+C feature space greatly limits its potential as the basis for compositional binning [80]. For example, for the TETRA method, a high z-score correlation (0.94) between two fragments indicates a probability of 79.5% that the two fragments originated from the same genome. Conversely, fragments that show absolutely no difference in G+C content only have a 10.4% chance of belonging to the same genome. Interestingly, it was noted that tetranucleotide frequency profiles are better able to distinguish between fragments at the species level than at the more general taxonomic ranks. For example, 99.5% of within-species fragment comparisons and 19.8% of between-species comparisons showed z-score correlations greater than the assignment threshold of 0.5, whereas only 22.3% of within domain (between-domain: 6.8%) comparisons exceeded a z-score correlation of 0.5.

Both TETRA and the G+C based method were also compared in their ability to successfully bin 6 fosmid-sized inserts from two low-complexity metagenomes shown to be involved in the anaerobic oxidation of methane [81]. The majority of the inserts

contained 16S rRNA genes, allowing the binning accuracy to be determined in the context of the accepted phylogenetic identities of the sequences. The G+C method succeeded in distinguishing between fragments from two genomes that had a moderately large difference in G+C of 10%, while it failed to distinguish between two genomes whose G+C contents differed by only 3.1%. Conversely, TETRA was able to bin all fragments correctly, exhibiting high within-genome z-score correlations of 0.82-0.91, and lower between-genome z-score correlations of <=0.60 in all cases.

Self-organizing Map (SOM)

Abe et al presented the application of a modified version of Kohonen's self organizing map [70] to the binning of metagenomic fragments, and demonstrated that such a method is capable of accurately binning short fragments into specific phylotypes based on similarities in tetranucleotide frequency profiles [45; 64]. Whereas many of the sequence-based classifiers previously discussed have ultimately relied upon a set of labelled training fragments in order to facilitate binning, the SOM approach is able to cluster DNA fragments into anonymous phylotypes based on similarities in tetranucleotide composition in a completely unsupervised fashion. In some instances, the resulting compositional bins have been shown to represent individual species or specific phylotypes, despite the fact that absolutely no taxonomic information or phylogenetic markers have been made available the classifier. Furthermore, after identifying a set of anonymous phylotypes using the SOM approach, these phylotypes may later be associated with known phylogenetic clades through a supervised SOM approach if a set of reference genomes is available.

In brief, the SOM is a form of artificial neural network, a machine learning method capable of mapping high dimensional data into a lower and often more comprehensible dimensional space while causing similar features to tend to be clustered within close proximity to one another in the resulting map. In terms of its algorithmic implementation, the SOM consists of a set of nodes referred to as neurons, each containing a weight vector of the same dimension as the feature space. Before features can be mapped to the SOM, the weight vectors of all neurons must first be initialized,

either by setting each of the weights to a small random value, or by assigning weights based upon a principal component analysis of the feature set. During the training phase of the SOM, features are sequentially projected onto cells within a 2D lattice or hexagonal grid of predetermined size, where each cell represents a specific SOM neuron. With each iteration, a single training feature is mapped to the SOM node with the most similar weight vector, and the set of neighboring nodes are updated in order to pull their respective weight vectors in the direction of the newly mapped feature. In this way, the SOM gradually assumes a topology in which similar features are clustered within local neighbourhoods in the resulting map. Once the SOM topology has been determined via the training process, and the weights of each of the SOM nodes have been defined, it is also possible to map additional features to the existing SOM without altering its topology. This optional mapping process can facilitate binning, by allowing for the association of each new training feature with a preexisting SOM neighborhood that was defined during the training phase. If a map is first constructed using a training set containing sequences of known taxonomic origin, then the subsequent mapping of anonymous sequences to the given SOM may facilitate binning to the associated phylogenetic clades, albeit in a supervised rather than unsupervised fashion. For the purpose of DNA classification, a given neighborhood of related features in a SOM may represent genomic fragments from a particular species or a more general phylotype.

Since SOMs are formed by projecting features onto nodes using a greedy assignment algorithm that continuously reorganizes the topology of the map, they are typically sensitive to the order of the input data. With this limitation in mind, Abe et al modified the standard SOM such that the topology of the resulting map remains consistent for any given set of training features, regardless of the order by which they are presented to the classifier. Additionally, an earlier study by Abe et al that relied upon non-symmetrized tetranucleotide frequency profiles indicated that species or phylotype clusters within a SOM were often subdivided into two smaller clusters based on the transcriptional polarity of the underlying DNA fragments. As it is difficult to determine the polarity of short DNA fragments from a metagenome, Abe et al extended their earlier study so that tetranucleotide frequency profiles were calculated across both strands of

each DNA fragment. The use of the resulting symmetrized tetranucleotide frequency profiles was shown to prevent the unnecessary sub-partitioning of phylotypes within the SOM, while maintaining clustering accuracy and reducing total computation time by close to 50%.

The performance of the SOM method was initially evaluated using 1kbp and 5kbp fragments from 81 completely sequenced prokaryotic genomes, representing 226Mbp of sequence in total. For each fragment length, a SOM was trained using all available fragments, and the resulting SOM topology was compared to the accepted taxonomic assignment of each of the fragments (based on their known genome of origin) in order to examine the SOM's binning accuracy. Overall, the 5kbp-trained SOM showed much higher binning potential than its 1kbp-based counterpart. For example, 74.6% of the 5kbp fragments were assigned to the correct species cluster, whereas only 40.6% of fragments were assigned correctly for the 1kbp-trained SOM. Interestingly, the percentage of correctly assigned 1kbp fragments nearly doubled when these fragments were mapped onto a SOM trained using 5kbp fragments, suggesting that even if the query fragments in a metagenome are relatively short, binning accuracy can be improved if the SOM is first trained using fragments that are longer than the query sequences. These results conflict with those of PhyloPythia, for which classification accuracy decreased in proportion to the difference in length between fragments (regardless of direction) in the training and testing sets [38].

In order to evaluate the binning accuracy of the SOM for real metagenomic sequences, Abe et al. next applied the SOM in order to classify sequences from the Sargasso sea metagenome [13]. A SOM was first constructed using 210,000 5kbp fragments from the 1502 known prokaryotes for which at least 10kbp of sequence was available. Next, 34,000 1kbp fragments were extracted from the nearly 4300 Sargasso metagenome contigs of at least 5kbp in length, and these fragments were subsequently mapped to the existing SOM in order to associate the fragments with known phylogenetic clades. The results of the mapping showed that the Sargasso fragments formed well-defined clusters in the SOM, and all of the known dominant members of the metagenome

were associated with these clusters. When the SOM mapping was repeated using 218,400 shorter 1kbp fragments extracted from the 134,600 metagenomic sequences >= 1kbp in length (and subsequently all 811,000 metagenomic sequences regardless of fragment length) the resulting clusters lacked the definition observed for the fragments derived from the longer contigs. This serves to highlight the influence of community structure on the expected resolution of phylogenetic classification. The fragments derived from the 5kbp or longer assembled contigs are expected to belong to the most abundant members of the Sargasso metagenome, and as such it is also expected that the fragments should form well defined clusters in the SOM. The shorter fragments, however, represent those sequences for which little to no read assembly was possible, and likely represent a multitude of flanking genomes that have much lower relative abundance in the metagenomic community, leading to the poorly defined clusters in the resulting SOM mapping.

In order to characterize complex metagenomic samples containing mixtures of prokaryotic as well as eukaryotic organisms, an essential aspect of any DNA classification system will be the ability to distinguish between the underlying prokaryotic and eukaryotic genomes present in the community. Abe et al evaluated the performance of the SOM method in this regard by constructing a SOM using the 210,000 5kbp fragments from the 1502 known prokaryotes for which adequate sequence exists, as well as 5kbp fragments from 6 fungi, 5 protozoa, and the zebrafish. The SOM showed remarkable accuracy in separating the fragments into eukaryotic and prokaryotic bins, assigning a mere 0.1% of the prokaryotic fragments into the eukaryotic clusters. When the Sargasso metagenome fragments >= 1kbp in length were subsequently mapped to the same SOM, the majority of fragments were assigned to the appropriate prokaryotic clusters, while 9.9% were assigned to various eukaryotic groups within the SOM. When these cross-domain mis-assignments were examined in detail, it was observed that the majority of the assignments fell within the clusters associated with unicellular eukaryotes, with few fragments assigned to the zebrafish cluster.

Semi-supervised Training of Classifiers

Seeded Growing Self-organizing Map (S-GSOM)

Chan et al created a novel semi-supervised metagenomic classifier based on an augmented version of the self-organizing map [36]. This method is particularly interesting because it is able to assign fragments into well-defined phylogenetic bins by automatically identifying sparse phylogenetic markers in a given metagenomic data set, and clustering the SOM nodes based on their affiliation with these markers. In essence, the S-GSOM method is quite similar to the previously discussed SOM method [64] in that both methods are capable of binning metagenomic fragments into anonymous phylogenetic bins based on their tetranucleotide profiles in the absence of a database of reference genomes. The S-GSOM classifier improves upon the basic SOM approach by implementing a more efficient SOM, the growing self-organizing map [82], and adding a semi-supervised post-processing step that utilizes sparse markers within the metagenome in order to improve the accuracy of the clustering of nodes within the SOM topology. Unlike the basic SOM binning approach, which may lead to clusters with ambiguous boundaries [45], the S-GSOM method aims to generate well-defined clusters while minimizing the assignment of ambiguous nodes for which no single best cluster assignment exists.

The key to the success of the S-GSOM approach is the post-processing step in which clusters are refined through the use of sparse phylogenetic seed sequences, in this case 16S rDNA flanking sequences. Chan et al opted to use 16S rDNA flanking sequences as the phylogenetic seeds in the cluster refinement algorithm because these sequences have already been shown to facilitate the binning of genomic contigs found in low-complexity metagenomes [83]. Additionally, 16S rDNA flanking sequences can easily be identified by their proximity to conserved rDNA sequences. Unlike 16S rDNA genes which are highly conserved between species and thus offer little signal for genome signature-based comparisons, the composition of 16S flanking sequences is much more variable in nature [84]. This increased variability in composition means that for a given genome, the flanking sequences are likely to exhibit similarities in compositional

characteristics with genomic fragments from the same genome. Conversely, because genome signature tends to vary more between genomes than within a genome [39; 43], the 16S flanking sequence from a given genome is likely to differ in composition from genomic fragments from an unrelated genome. In the case of the S-GSOM, the seed sequences are initially combined with the set of 8kbp genomic fragments to be classified by the SOM, and compositional similarities between the test fragments and these labelled seed sequences are used to determine the cluster assignments in the post-processing step.

The binning accuracy of the S-GSOM method was compared against three other binning methods, namely PhyloPythia [38], Chi-Squared [25], and Blast distribution [25], in their respective abilities to bin both the Phrap- and Arachne-generated assemblies of the low complexity (simLC) and medium complexity (simMC) simulated metagenomic datasets from the FAMeS study [25]. To quantify the binning accuracy in each case, Chan et al calculated the total percentage of binned contigs, the sensitivity, and the specificity of each method at the taxonomic ranks of class, order, and family. The accuracy scores were evaluated using two subsets of contigs from each of the simulated datasets: 1) the subset of contigs of at least 8kbp in length, and 2) the subset of contigs consisting of 10 or more reads.

Overall, the S-GSOM method exhibited reasonable accuracy scores for both metagenomic datasets using each of the two subsets of contigs, outperforming the Chi-Squared and BLAST distr methods in all cases. At the taxonomic rank of family, the most specific rank examined, the S-GSOM outperformed PhyloPythia for all cases except the subset of simLC contigs >= 8kbp in length. For this exceptional case, the S-GSOM achieved both lower sensitivity (PhyloPythia: 95% vs. S-GSOM: 89.1%, Arachne assembly) and specificity (PhyloPythia: 95% vs. S-GSOM: 89.1%, Arachne assembly), although PhyloPythia was able to bin approximately 6% more contigs than the S-GSOM for this dataset. Interestingly, for the subset of >8bkp fragments from the Arachne assembly of the more complex simMC metagenome, the S-GSOM greatly outperformed PhyloPythia, binning nearly twice as many contigs (92.69% vs. 47.51%), and achieving both a higher sensitivity (89% vs. 40.1%) and specificity (92.7% vs. 47.5%). At the less

specific taxonomic ranks of class and order, the S-GSOM and PhyloPythia generally demonstrated comparable sensitivities and specificities, although PhyloPythia was typically able to bin a higher total percentage of contigs. PhyloPythia's increased performance at more general ranks may be due to the fact that it bases its classification on both pentanucleotide and hexanucleotide frequency profiles of the query fragments, whereas the S-GSOM relies solely on tetranucleotide frequencies. As reported in the PhyloPythia manuscript, longer oligonucleotides are better able to model the compositional signatures of the more general taxonomic ranks, which may give PhyloPythia an advantage when its performance is compared to that of other classifiers at any of the less specific taxonomic ranks.

CompostBin

CompostBin is a semi-supervised metagenomic binning system, allowing for the accurate binning of single ~1000bp reads from simulated low to medium complexity metagenomes [65]. As opposed to several of the other sequence-based methods that rely on machine learning methods in order to classify DNA fragments, CompostBin instead combines a novel principal component analysis (PCA) technique with a semi-supervised clustering algorithm in order to facilitate classification of fragments based on their hexanucleotide frequency profiles. For DNA fragments in a metagenome, the hexanucleotide frequency profiles from each fragment are first projected into a lower dimensional space using a weighted PCA technique, and features within this lower dimensional space are subsequently partitioned into taxonomic bins using a normalized cut clustering algorithm. Although the PCA component of CompostBin is unsupervised in nature, the normalized cut algorithm is largely dependent on outside information in order to facilitate taxonomic binning. Notably, the normalized cut algorithm requires input relating to both the number of taxonomic bins present in the dataset, as well as the presence of known phylogenetic markers on specific reads in the metagenome.

PCA [85] is a multivariate data analysis technique that is often used to reduce the dimensionality of datasets by identifying the set of features (principal components) that contribute the greatest influence toward the variance of the data. When applied to the

binning of metagenomic fragments based on their oligonucleotide frequency profiles, the goal of a PCA is to identify the oligonucleotide patterns that best describe the compositional variation between the taxonomic classes present within the metagenome. The authors of CompostBin noted that in the likely case that the relative proportions of the individual members of a metagenomic community are unbalanced, then traditional PCA might simply identify the principal components that describe the within-genome variation in the predominant genome(s), rather than the components that capture the between-genome compositional variation that would ultimately facilitate taxonomic binning. As such, the authors devised a weighted PCA algorithm, whereby each fragment in a metagenome receives a weight inversely proportional to the relative abundance of that read in the dataset. The weighted PCA algorithm then takes these weights into account when identifying the principal components, by decreasing the influence of each fragment in relation to its abundance within the dataset. By applying this weighted PCA technique using the complete set of hexanucleotide frequency profiles associated with a metagenome, CompostBin is thus able to reduce the feature space from the 4096 possible hexanucleotides to the 3 most influential principal components specific to the given dataset.

Once the hexanucleotide frequency profiles have been transformed via weighted PCA, CompostBin next applies the normalized cut clustering (NCC) algorithm to partition the features into the relevant taxonomic bins. In order to successfully partition the feature space, the algorithm requires that a portion of the features contain labels that associate these features with known phylogenetic clades. Additionally, the algorithm must be informed of the number of taxonomic bins that are present in the metagenome. The NCC utilizes a weighted graph representation of the 3-dimensional feature space, where features are represented as nodes in the graph, and the vertices connecting each node are weighted in relation to each feature's association with one or more of the labelled phylogenetic markers. For each iteration of the NCC algorithm, the set of vertices in the graph are bisected into two subsets such that the weights connecting the vertices within each cluster are maximized, while the weights connecting the vertices between subsets are minimized. NCC is applied recursively to the resulting subsets of vertices in order to

achieve the desired number of taxonomic bins.

In order to evaluate the binning accuracy of CompostBin, the authors created 12 simulated metagenomes of varying complexity. Each simulated metagenome contained 2-6 genomes in relative proportions ranging from 1:1 to 1:14. Care was taken to ensure that the 12 metagenomes contained a variety of community structures as well as a range of phylogenetic and compositional diversity. For each metagenome, sequencing reads with an average length of 1000bp were simulated from the component genomes using the ReadSim package and compiled using the appropriate proportions [77]. In addition to the simulated metagenomes, CompostBin was also evaluated in its ability to correctly bin sequencing reads from the glassy-winged sharpshooter metagenome [86] into the two predominant bacterial species previously identified using a phylogenetic marker approach.

The performance of CompostBin for each metagenome was reported in terms of the class-normalized error rate (i.e., corrected for the number of instances from each taxonomic group), where the individual class-level error rates were determined for each genome in the given dataset, and the class-normalized error rate was then calculated as the average of all class-level error rates.

CompostBin exhibited low class-normalized error rates across all simulated metagenomic samples, ranging from 0.28% to 10%. The lowest observed error rate of 0.28% was achieved for a low-complexity metagenome consisting of *Thermofilum pendens* and *Pyrobaculum aerophilum* in a 1:1 ratio. Interestingly, a low-complexity metagenome containing two organisms that differ at the taxonomic level of genus, *Escherichia coli* and *Yesinia pestis*, showed the highest error rate at 10%. For the metagenomes containing reads from 3-6 individual genomes, the error rates varied from 1.96% - 7.7%. The sole metagenome comparing two species of the same genus, *Bacillus halodurans* vs. *Bacillus subtilus* in a 1:1 ratio, showed an error rate of 6.48%. CompostBin also performed with comparable accuracy when faced with the glassy-winged sharpshooter metagenome, classifying the metagenomic fragments with an error rate of 9.04%.

Limitations of the Existing Methods

Although a variety of methods have shown great promise in their ability to classify metagenomic DNA fragments, the performance of such methods is often heavily dependent on several factors that are not easily controlled, such as the length of the DNA fragments in the sample, the complexity of the given metagenomic community, compositional similarities between members of the community, and the existence of closely related sequences within the various reference databases. For example, sequence-composition-based approaches (PhyloPythia, TACOA) and homology-based approaches (BLAST distr, CARMA) perform best when applied to moderate-length fragments from low-complexity metagenomes (i.e, communities comprising a small number of well-represented organisms) for which the predominant members have close relatives in the respective sequence databases or training sets. All of these methods suffer a drastic decrease in performance when attempting to classify shorter sequences, fragments from complex metagenomic communities, or sequences for which a close relative is not available for comparison. Unsupervised methods such as the various SOM clustering approaches do not explicitly depend upon reference databases of known sequences, however they tend to succeed only in binning longer fragments while the resulting taxonomic clusters are often poorly defined and the assignment of fragments into the existing phylogenetic hierarchy is not possible without performing a comparison against a reference database. Even the semi-supervised methods such as CompostBin and S-GSOM break down when attempting to classify complex communities or in cases where identifiable markers are absent from the metagenomic dataset. Although it is expected that the performance of existing methods will gradually improve as sequencing technologies allow for longer read lengths and reference databases become more representative of true microbial diversity, it is likely that the performance of such methods will still suffer when faced with complex metagenomic communities or even simple communities that contain a number of compositionally similar organisms. If the phylogenetic composition of these communities is ever to be understood, it is of the utmost importance to identify controllable factors that may influence classification, and attempt to leverage these factors in order to improve the classification accuracies of the

existing methods. Additionally, if there are fundamental limitations to the sequence-composition and sequence-homology approaches to DNA classification, the characterization of these limitations may help us to understand the 'best case' classification accuracies that we may expect for a given community.

To date, most classification methods report global accuracy scores at the various taxonomic ranks without paying particular attention to the individual comparisons that may potentially skew the overall performance of the classifiers. By performing pairwise classification as opposed to multiclass classification, we may be able to bring attention to specific pairs of genomes that are easier or more difficult to classify than might be expected. Closer examination of such pairs of genomes may even suggest mechanisms by which existing classifiers may be improved. Pairwise genome classification will be a fundamental aspect of the experiments outlined in both Chapters 3 and 4.

All of the existing methods for DNA classification show a trend of decreasing classification accuracy in proportion to an increasing level of specificity of the taxonomic rank at which sequences are being compared. This is to be expected, as in general, two organisms that have a close phylogenetic relationship are likely to have similar genome signatures, which will in turn reduce distinguishability. None of the existing methods, however, provide a clear understanding of exactly how classification accuracy varies in relation to factors such as the level of conservation of orthologous sequences for a pair of genomes, differences in G+C composition, genomic similarity based on shared loci or conserved marker genes, and tetranucleotide composition. An in-depth analysis may help us to understand the bounds of classification imposed by such measures of genome similarity, and perhaps allow us to identify outlier genome comparisons that provide additional insight into the classification problem. These features and more will be examined in Chapters 3 and 4.

In cases where 100% classification accuracy is not achieved for a given pair of genomes, it is important to understand the factors that contribute to the decrease in distinguishability. Obvious confounding factors may include recent LGT events and the presence of phage DNA or pathogencity islands in the pair of genomes. In many

instances, such sequences may essentially be indistinguishable because nearly identical sequences exist in both genomes. Other factors may provide more fruitful avenues for the improvement of existing classifiers. For example, Chan et al reported that sequence chimerism had an immense impact on classification accuracy for their classifier, and even avoided attempting to classify sequences <= 8kbp in order to reduce its impact on the S-GSOM method [36]. Although this type of sequence heterogeneity referred specifically to chimeric contigs containing sequence from multiple genomes, it may suggest that a more generalized concept of coding vs. non-coding sequence heterogeneity within individual sequence reads may also influence classification. Similarly, recombination might result in the presence of multiple phylogenetic signals within a single read. Differences in the relative conservation of certain classes of proteins may impact the classification of DNA fragments containing sequence derived from these different protein classes [38]. For instance, fragments of genes encoding highly conserved ribosomal proteins may be much harder to distinguish on the basis of genome signature than genes encoding less conserved metabolic pathways. Additionally, in some cases factors such as habitat or lifestyle may lead to the convergence of genome signature for specific pairs of organisms [87; 88], causing an otherwise unexpected decrease in distinguishability for a pair of genomes. Conversely, closely related organisms that have undergone rapid evolution may in fact exhibit increased distinguishability in comparison to what might be expected based upon a phylogenetic marker gene approach. All of these potential confounding factors will be examined in depth in Chapter 4.

It is widely accepted that the the relative frequencies of specific oligonucleotide patterns can be utilized to capture genome signature and distinguish between genomes that exhibit sufficient differences in composition. In reviewing the various sequence-composition based classifiers, it is evident that there is no single best set of parameters for capturing genome signature using this oligonucleotide frequency approach. Some methods, such as the Naïve Bayes classifiers, report the highest classification accuracies while using frequencies of long 9-15 nt oligonucleotides (although the results reported for the longest k-mers might be artifacts as discussed in the section outlining the NBC method), while others report the best performance while using tetramer [40],

pentanucleotide [38], or hexanucleotide frequencies [38; 65]. To further complicate the issue, Bohlin et al even suggested that little signal is gained by using oligonucleotide patterns longer than 6 nt, in stark contrast to the results presented in the Naïve Bayes studies [52; 62; 89]. Furthermore, certain classifiers such as PhyloPythia and the Chi-Squared classifier make use of degenerate oligonucleotide patterns (i.e., the classifier may use hexanucleotide patterns that contain one or more IUPAC 'N' characters, allowing for relaxed matching of each hexanucleotide and thus decreasing the sparsity of the resulting feature vector), claiming increased performance over strict oligonucleotide patterns. The inconsistency of optimal parameters within the literature justifies an examination of the impacts of oligonucleotide pattern length and the degeneracy of patterns on classification accuracy. Such a study will be presented in Chapter 2.

Chapter 2 – Investigating the Influences of DNA Recoding, K-mer Size, and DNA Fragment Length on Classification Accuracy

Motivation

Existing DNA classification systems such as PhyloPythia [38], CompostBin [65], and TETRA [63] have typically utilized the relative frequencies of short oligonucleotides (k-mers) as a means of quantifying genome signature. Although these methods have demonstrated that it is possible to distinguish between genomes on the basis of their k-mer frequency profiles, the feature space sizes associated with k-mer frequency data sets impose restrictions on their application to classification. Given the 4-letter nucleotide alphabet, a k-mer will result in a feature space of 4^k elements, leading to very large feature spaces for even relatively small values of k (ex: $4^8 = 65536$ features). Large feature spaces can lead to prohibitive computational and memory requirements, and can also reduce the performance of machine learning and statistical methods that are susceptible to the "curse of dimensionality".

Various DNA recoding schemes have been used in order to overcome compositional biases in genome sequences or to transform such sequences so that they may be analyzed using advanced signal processing techniques. One common DNA recoding scheme transforms a given genome sequence into 4 binary sequences in which 1s are used to denote the presence (0's the absence) of one of the four possible nucleotides {A, C, G, T} present in the source genome. This binary recoding scheme has been used in order to apply wavelet transform techniques to genome sequences [90; 91] and to investigate the fractal nature of DNA [92]. Similarly, binary recoding was used by Hill et al. [93] in order to apply Chaos Game Theory to the visualization of genome sequences. Binary DNA recoding schemes were also used in an attempt to identify questionably aligned genome sequences [94].

A second DNA recoding scheme, RY-recoding, removes G+C compositional biases in DNA by generalizing such sequences so that they contain only the symbols for

purine (R) and pyrimidine (Y) bases. Phillips et al. demonstrated that RY-recoding mitochondrial DNA sequences prior to phylogenetic analyses served to both reduce compositional biases and enhance the phylogenetic signal [95].

The current study employs several DNA recoding schemes in an attempt to reduce the number of features associated with a particular k-mer length. For each recoding scheme, DNA sequences are first mapped to binary sequences based on criteria such as nucleotide identity and purine/pyrimidine content (Table 2.1). The resulting binary sequences are subsequently analyzed in order to determine the frequencies of specific binary patterns of various lengths, and these frequency profiles are used to train multiclass support vector machine (SVM) based classifiers. Aside from the reduced memory requirements, it is anticipated that binary-recoded DNA sequences will perform at least as well as non-recoded DNA for the purpose of sequence classification using SVMs.

Two types of binary SVM classifiers are presented here: a s*imple binary classifier*, and a *combined-binary classifier*. The simple binary classification system recodes DNA by assigning matching nucleotides (or classes of nucleotides) the value of 1 and all other nucleotides the value of 0 in a given DNA sequence, and then uses the observed frequencies of binary patterns as input for constructing SVM models. The combined-binary system recodes the given DNA sequence individually using the simple binary recoding scheme for each of the nucleotides {A, C, G, T} and then uses the combined set of frequencies of the binary patterns for each of the resulting 4 recoded sequences to construct SVM models. Although the combined-binary classifier requires 4 times the feature space of the simple binary classifier, it nonetheless requires considerably fewer features than the plain k-mer-based classifier for pattern lengths greater than or equal to 3 nt (see Table 2.2).

Table 2.1: Description of DNA Recoding Schemes

Recoding Scheme	Recoding criteria
A	Each A in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.
C	Each C in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.
G	Each G in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.
T	Each T in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.
AT	Each A or T in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.
CG	Each C or G in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.
AG	Each purine in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.
CT	Each pyrimidine in the nucleotide sequence is recoded to a 1. All other nucleotides are recoded as 0.

Table 2.2: Comparison of the Number of Features per SVM Training Instance for Each Classifier Type.

Values in bold indicate the largest SVM feature space examined in this set of experiments for each type of classifier.

Pattern Length (n)	**Simple binary classifier (2^n features)**	**Combined-binary classifier ($4*2^n$ features)**	**K-mer classifier (4^n features)**
1	2	8	4
2	4	16	16
3	8	32	64
4	16	64	256
5	32	128	1024
6	64	256	**4096**
7	128	512	16384
8	256	1024	65536
9	**512**	**2048**	262144

The primary goal of this study is to evaluate the performance of SVM-based DNA classification systems using several DNA recoding schemes, and to compare the results against SVM classifiers in which no DNA recoding is used. Additionally, the influences of k-mer size and DNA fragment length on classification accuracy will be investigated both with and without the use of DNA recoding.

Support Vector Machines

The support vector machine (SVM) is a state-of-the-art machine learning method that has been successfully applied to a range of classification problems, including speech recognition [96], image recognition [97], microarray expression profiling [98], and text classification [99]. When presented with a set of training data consisting of labelled features spread across multiple classes, the support vector machine constructs a model by identifying an appropriate set of hyperplanes that partition the feature space into training classes based on the class labels. Hyperplanes are selected such that the margins between the boundary features (referred to as support vectors) within each class are maximized, and thus the SVM is referred to as a maximum margin classifier. In cases where the feature sets belonging to two or more classes overlap in the feature space, implying that perfect class distinction is not possible, the SVM chooses appropriate hyperplanes by minimizing an error function related to the number of features that are incorrectly partitioned. This error function depends upon a cost parameter C that determines the error penalty associated with each misclassified feature. Since larger cost parameters are associated with larger error penalties, choosing too large a cost parameter may result in overfitting of the model. Conversely, choosing too small a cost parameter will result in a model that is overly permissive to misclassifications. This cost parameter is dataset specific, and heuristic grid searches are often used in order to identify appropriate values of C.

As with the modified k-NN algorithm (see TACOA classifier in Chapter 1), the SVM incorporates the use of kernel functions in order to alleviate the effects of the curse of dimensionality. Common kernel functions implemented in SVMs include linear, Gaussian, polynomial, and sigmoidal functions [100], and the relative performance of the

kernel functions has been shown to vary depending on the underlying classification problem at hand. For example, the linear kernel has been shown to outperform the other kernels when applied to text classification (http://www.csie.ntu.edu.tw/~cjlin/liblinear/), whereas the gaussian kernel has been shown to be most appropriate when applied to the classification of DNA oligonucleotide frequency profiles [38]. In order to use the Gaussian kernel function in conjunction with the SVM, the kernel width parameter γ must be passed to the SVM algorithm during the training phase. Like the cost parameter C discussed above, γ is dataset-dependent, and grid searches are frequently used to choose reasonable values of γ.

Although the support vector machine may be applied to multiclass problems, the core SVM algorithm is a binary classifier. In order to perform multiclass classification, available SVM implementations transparently reduce an n-class problem into a set of $\binom{n}{2}$ one-against-one [100] or one-against-the-rest (http://pyml.sourceforge.net/) binary classifiers, and ultimately use a voting procedure in order to aggregate the results from the individual binary classifiers into a multiclass prediction. In the case of libSVM [100], each feature in an n-class problem is evaluated using $\binom{n}{2}$ one-against-one SVMs, and the given feature is predicted to belong to the class with the highest number of votes produced by the complete set of pairwise classifiers. In the event of a tie, the feature is predicted to belong to the class with the lowest numerical ID.

Experimental Design

DNA Recoding Schemes

Although PhyloPythia's use of k-mer frequency profiles has been shown to provide very accurate classification of DNA fragments [38], this approach produces high-dimensional SVM training and test sets. The use of k-mers of the 4 nucleotides {A, C, T, G } quickly results in an enormous feature space of size 4^k, which increases the size of the SVM training files, the associated memory requirements, and the computational effort required in order to construct and utilize the resulting SVM models. Rather than

focusing on the frequencies of k-mers, the experiments in this chapter aim to evaluate the ability of binary patterns of recoded nucleotides to capture the genome signature exhibited by DNA fragments.

Table 2.2 lists the various DNA recoding schemes that are used in this set of experiments. Each of the recoding schemes is used individually in order to create 8 simple binary classifiers (Figure 2.1a,b). Additionally, a single combined-binary classifier is built using the combined set of frequencies from the A, C, T, and G recoding schemes.

Data Acquisition

The procedures in this set of experiments make use of a set of 10 completely sequenced Bacterial and Archaeal genomes (see Table 2.3). Genomes were selected to ensure that both closely related and distantly related organisms were represented. The complete DNA sequences and all associated information for the organisms in Table 2.3 was obtained from the Joint Genome Institute IMG/M online service [61] on October 12th, 2007.

Figure 2.1. Binary Recoding and Parameterization of a DNA Sequence.

a) Recoding of the given DNA sequence using the simple binary recoding scheme for adenine "A". b) Purine recoding of the source DNA sequence. c) Parameterization of a purine recoded DNA sequence for a pattern length of 3 nt. The total counts of all overlapping 3-mers are first tallied and subsequently divided by the fragment length in order to determine the 3-mer frequency vector.

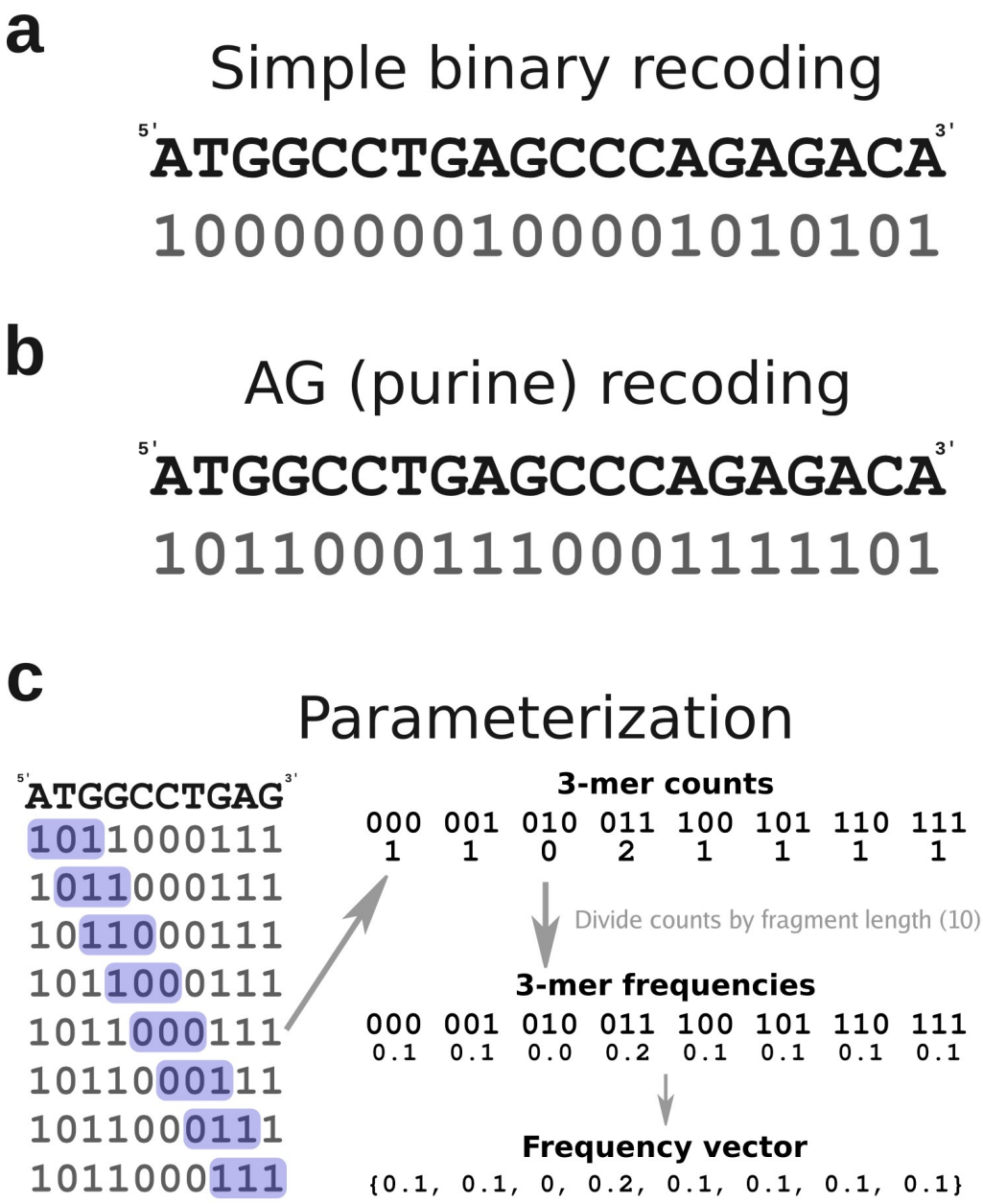


Table 2.3: List of Genomes Selected for use in the Experimental Procedures.

Organism Name	NCBI Accession #	Domain	Phylum	Class	Genome Size (bp)
Acidobacteria bacterium Ellin345	NC_008009	Bacteria	Acidobacteria	Acidobacteria	5650368
Bacillus anthracis str. Ames	NC_003997	Bacteria	Firmicutes	Bacilli	5227293
Bacillus cereus E33L	NC_006274	Bacteria	Firmicutes	Bacilli	5843235
Dechloromonas aromatica RCB	NC_007298	Bacteria	Proteobacteria	Betaproteobacteria	4501104
Escherichia coli O157:H7 str. Sakai	NC_002695	Bacteria	Proteobacteria	Gammaproteobacteria	5594477
Halobacterium sp. NRC-1	AE004437	Archaea	Euryarchaeota	Halobacteria	2571010
Legionella pneumophila str. Lens	NC_006369	Bacteria	Proteobacteria	Gammaproteobacteria	3405519
Methanococcus maripaludis strain S2	BX950229	Archaea	Euryarchaeota	Methanococci	1661137
Staphylococcus aureus subsp. aureus USA300	NC_007793	Bacteria	Firmicutes	Bacilli	2917469
Synechococcus sp. JA-3-3Ab	NC_007775	Bacteria	Cyanobacteria	-	2932766

All DNA sequences were downloaded as flat text files in the FASTA format. Each text file contained the full genome sequence of each organism, including multiple chromosomes and/or plasmids, where applicable.

Genome Parameterization

The nucleotide sequence of each genome was completely partitioned into non-overlapping fragments of size 500 nt, 1000 nt, and 5000 nt, resulting in three sets of fragments for each genome. For genomes that contained more than one DNA molecule (plasmids or multiple chromosomes), all associated DNA sequences were concatenated to produce one large sequence prior to partitioning the genome into fragments.

For each genome, 3000 DNA fragments from the previous partitioning step were randomly selected and recoded using the 8 DNA recoding schemes listed in Table 2.2. Each of the 3000 fragments was scanned along the coding strand from beginning to end using a sliding window approach (window size = k, window step = 1) in order to determine the total counts of all possible 2^k binary patterns present in the fragment, for k ε {3, 4, 5, 6, 7, 8, 9} (Figure 2.1c). Additionally, the total counts of all possible 4^k k-mer patterns of nucleotides were also tabulated for each of the 3000 fragments using the same approach, with k ε {3, 4, 5, 6}. In all cases, pattern counts were converted to frequencies by dividing each count by the fragment length. The resulting frequency vectors were subsequently scaled between -1 and 1 using the scale.py script from the libSVM package [100]. Frequency files were then split evenly to produce SVM training and testing files each 1500 instances in length, for each combination of fragment length, pattern length, and classifier. libSVM's subset.py script was used to split the frequency files in a stratified fashion, such that the frequency of each class was identical in both the training and test files.

Building and Evaluating SVM Models

Training the SVMs

For each SVM training file, a grid search was performed in order to determine

reasonable values for C and γ. In each case, 300 training instances were used to perform a grid search using libSVM's grid.py script with 5-fold cross validation. An SVM model was then built by running the program 'svm-train' on the 1500 instance training file using the C and γ values previously determined in the grid search. The Radial Basis Function (RBF) kernel was used in both the grid searches and the training of the SVMs, as this has previously been shown to outperform other kernel functions in a similar implementation of DNA sequence classification [38]. The perl module Time::HiRes v1.20 (http://search.cpan.org/~deweg/Time-HiRes-01.20/) was used to record high-resolution timestamps immediately before and immediately following the execution of svm-train. These timestamps were used to determine the total training time required for each SVM.

An additional set of frequency files was prepared for both the combined-binary classifier and the k-mer classifier. In this second set of frequency files, the amount of training sequence was fixed at 600,000 nt for each of the models, by varying the numbers of fragments in each training set depending on the fragment length being examined. For fragments of length 500 nt, 1200 fragments were evaluated. For fragments of length 1000 nt, 600 fragments were evaluated. And lastly, for fragments of length 5000 nt, 120 fragments were evaluated. The purpose of varying the number of fragments in relation to fragment length was to test whether or not the observed increased performance of the SVM classifiers for large fragment sizes was due to increased training sequence relative to the shorter fragments (i.e., in the original trials, the training sets always consisted of 3000 fragments regardless of fragment size). For each test case, 300 instances were used to perform grid searches (see above), and SVM models were built using half of the available frequency profile data.

Testing the SVMs

For each encoding strategy, each corresponding SVM model was used by the 'svm-predict' program to classify fragments from the test files that had the same pattern length as the training file used to build the model. Although pattern length remained consistent between the training and testing file involved in each comparison, separate SVM runs were used to evaluate all possible training fragment length and test fragment

length combinations. Time::HiRes was used to calculate the running time of svm-predict in the same manner as it was applied to svm-train above. Average sensitivity and specificity values were calculated from the output of svm-predict, and averaged over the three trials.

Evaluating SVM Performance

For each test run of a given SVM, the input test file and the resulting prediction file were compared in order to calculate the average sensitivity and average specificity of the given SVM. For each genome in the test/prediction files, sensitivity was calculated as:

$S_n = \frac{TP}{(TP+FN)}$ where TP represents true positives, and FN represents false negatives.

The average classification sensitivity was then calculated as the average of all of the class-level sensitivities. Likewise, specificity was calculated as:

$S_p = \frac{TN}{(FP+TN)}$ where TN represents true negatives and FP represents false positives.

As above, the average classification specificity was calculated as the average of all class-level specificities.

Results

Comparison of Classification Sensitivities of all Classifiers

For each of the classifiers, the classification sensitivity was examined over fragments of length 500, 1000, and 5000 nt. In the case of the binary classifiers, patterns of length 3-9 nt were examined, whereas patterns of length 3-6 nt were chosen for the k-mer classifier in order to maintain reasonable training and testing times for the SVMs. Sensitivity was calculated for each combination of classifier, fragment length, and pattern length as the average over 3 replicate trials.

As can be seen in Figure 2.2, the k-mer classifier (max Sn = 88.5%) generally

outperforms the combined-binary classifier (max Sn = 86.5%), which always outperforms the simple-binary classifiers for all patterns tested (max Sn = 81.4%). All classifiers exhibited a general trend of increasing sensitivity in proportion to fragment length. The range of pattern lengths examined appears to convey comparable sensitivity for each combination of classifier and fragment length.

The highest sensitivity (88.5%) was achieved by the k-mer classifier using a fragment length of 5000 nt and pattern lengths of both 3 and 5 nt. The lowest sensitivity observed was 32.7%, by the simple binary classifier using the 'T' simple binary recoding scheme.

Of all of the binary classifiers, the combined-binary classifier offered the sensitivity (60.7% - 86.5%) most similar to that of the k-mer classifier. Although there is a large discrepancy between sensitivities of the various classifiers for small fragment sizes, the overall difference in classification sensitivity decreases as the fragment size increases, demonstrating that even the worst of the simple binary classifiers is able to capture the genome signature for longer fragments.

Among the various binary recoding schemes examined, specific recoding schemes and their reverse complements achieve very similar sensitivities. For example, the “A” and “T” lines track together, as do { C, G } and { AG, CT }. Although the self-palindromes AT and CG are not reverse complements of one another, their sensitivities are also nearly identical across all pattern and fragment length combinations.

The apparent decrease in sensitivity of the k-mer classifier for the 5000 nt fragments with pattern lengths of 4 nt and 6 nt is an artefact that can in each case be attributed to one replicate trial (of three) where the classifier performed inconsistently relative to the other two replicates. The standard deviations between the three replicates for these two pattern lengths are 0.181 for k=4 and 0.068 for k=6, compared to an average standard deviation of 0.010 (range 0.002 – 0.027) across all other k-mer pattern length and fragment length combinations. These artifacts can likely be attributed to a grid search performed on an unrepresentative subset of the training data, leading to the selection of C and γ values that perform poorly when applied to the entire training set. If

the inconsistent results are excluded, the average sensitivities are 89.1% for k=4, and 87.4% for k=6, resulting in a much smoother line for the sensitivities of the 5000 nt fragments. Increasing the number of items used in the grid searches might have avoided these inconsistencies, with the trade-off of increased running time.

Comparison of Combined-binary Classifier vs. K-mer Classifier Using Fixed Genome Coverage

Classification sensitivity generally increases in proportion to length of the fragments used to build and test each SVM (Figure 2.2). It should be noted that this increase might be caused by the fact that for each of the fragment lengths tested (500 nt, 1000 nt, 5000 nt), 1500 fragments of each size were used to build each associated SVM. In essence, the SVMs built with the larger fragment sizes had an advantage in that they had been exposed to a much larger portion of each of the genomes than the SVMs built from the smaller fragment sizes. In order to determine whether or not this difference in coverage was responsible for the apparent increase in sensitivity with fragment size, a second set of SVMs was built using different numbers of fragments depending on the fragment sizes. 'Fixed coverage' versions of both the combined-binary classifiers and the k-mer classifiers were created, using fragment sizes of 500 nt (1200 fragments), 1000 nt (600 fragments), and 5000 nt (120 fragments).

After correcting for potential bias due to differences in genome coverage, the general increase in classification sensitivity with fragment size is still apparent (Figure 2.3), indicating the higher accuracy associated with larger fragment sizes is not due to increased genome coverage. The k-mer classifier achieved the highest sensitivity in this experiment (89.1%) using a pattern length of 5 nt and a fragment length of 5000 nt.

Figure 2.2: Comparison of Average Classification Sensitivity Over Varying Fragment and Pattern Length Combinations for all Classifiers.

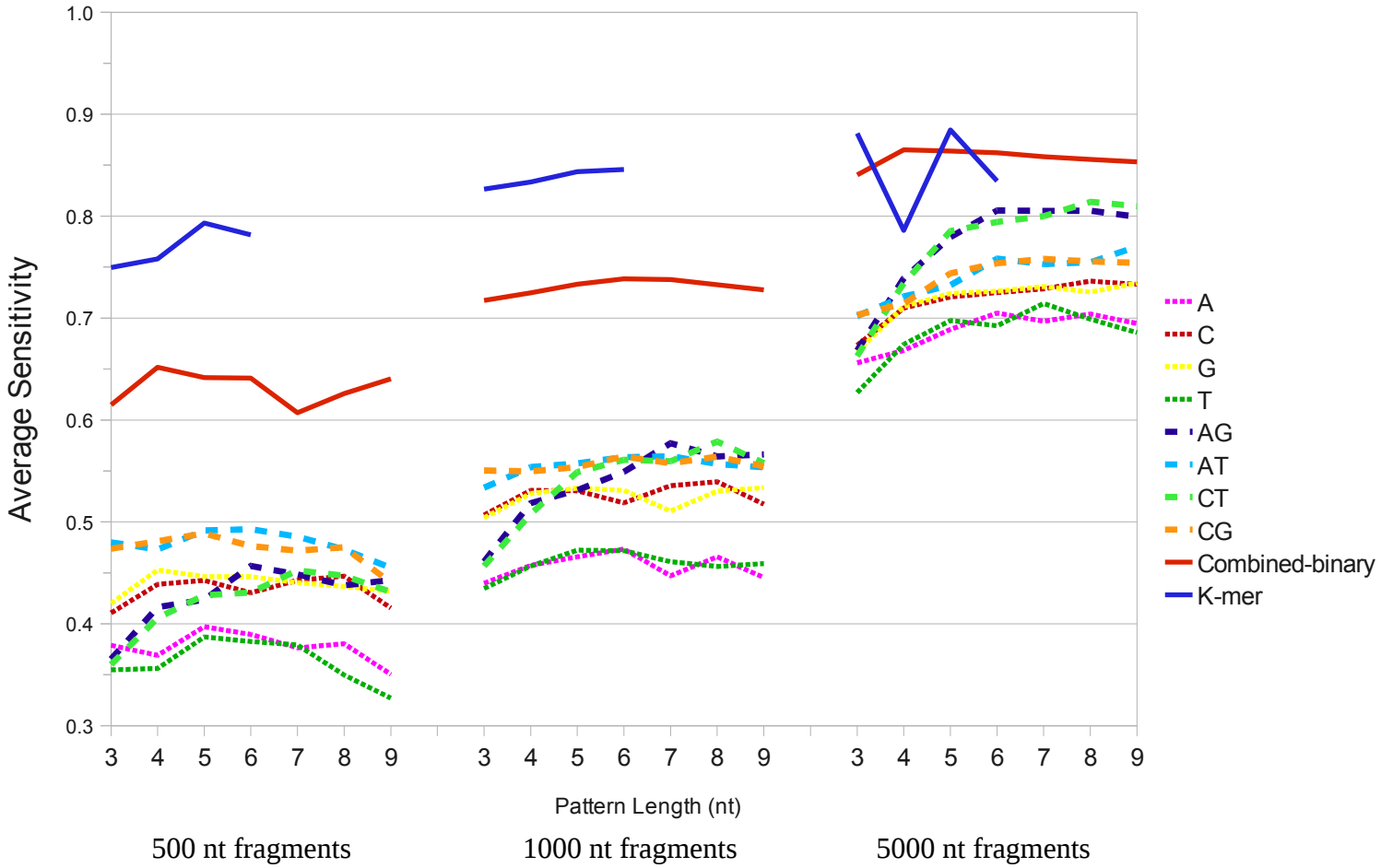
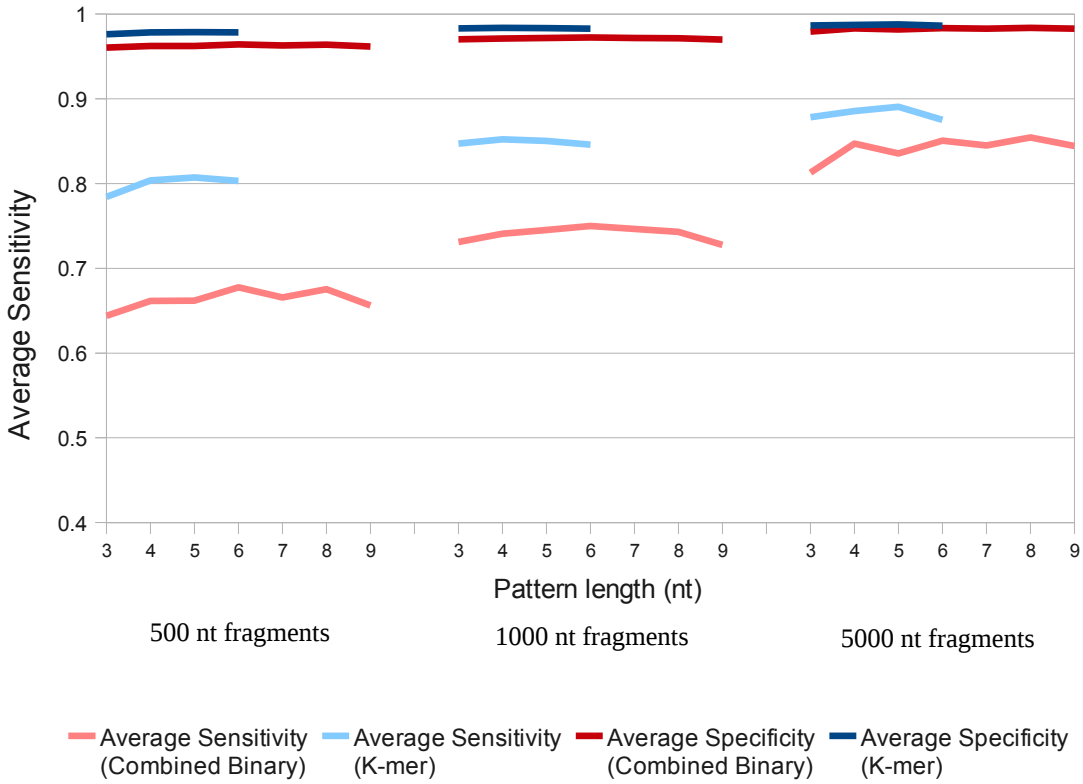


Figure 2.3: Average Sensitivity and Specificity Over Varying Fragment and Pattern Length Combinations for the Combined-binary and K-mer Classifiers.



The highest sensitivity achieved by the combined-binary classifier was 85.4% for a 7 nt pattern length and 5000 nt fragment length. Specificities for the combined-binary classifier and k-mer classifiers ranged from 96.0%-98.4% and 97.6%-98.8% respectively.

Comparison of the Classification Sensitivities of the Combined-binary and K-mer Classifiers for Test Fragments of Different Size Than Those Used to Build the Classifiers

For the combined-binary and k-mer classifiers, each SVM model trained for a given fragment length was used to classify all test sets across the full range of fragment lengths (500 nt, 1000 nt, and 5000 nt). This cross-testing of models and test sets was performed in order to judge the classifiers' ability to generalize and classify fragments that were not necessarily the same size as those used to train the classifier.

In examining Figures 2.4, 2.5, and 2.6 a few things are readily observable. First and foremost, the k-mer classifier outperformed the combined-binary classifier in almost all cases (except for 2 points in the comparison of average sensitivity vs. pattern length for models trained using 1000 nt fragments). Although the combined-binary classifier offered comparable classification specificity in some cases, the k-mer classifier provided better sensitivity, particularly with models built from the smaller fragment sizes. For example, in Figure 2.6, the combined-binary classifier achieved a maximum sensitivity of 86.5% when the 5000 nt trained model was tested against 5000 nt fragments using a pattern length of 4. This compares quite favourably to the k-mer classifier's performance using 5000 nt trained models (max sensitivity = 88.5%). For models trained using shorter fragment sizes, however, the difference in sensitivities between the two classifiers increases dramatically. For the 500 nt trained models, the combined-binary classifier achieved a maximum sensitivity of only 74.2% (for the 5000 nt fragments), whereas the k-mer classifier had a maximum sensitivity of 86.9%.

Figure 2.4: Average Sensitivity vs. Pattern Length for SVM Models Trained Using 500 nt Fragments

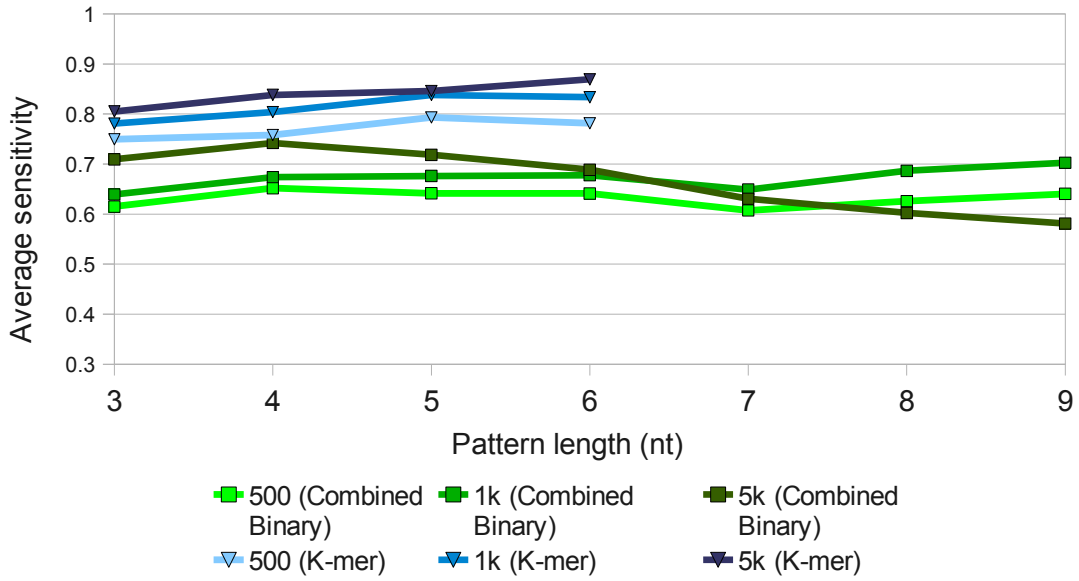


Figure 2.5: Average Sensitivity vs. Pattern Length for SVM Models Trained Using 1000 nt Fragments

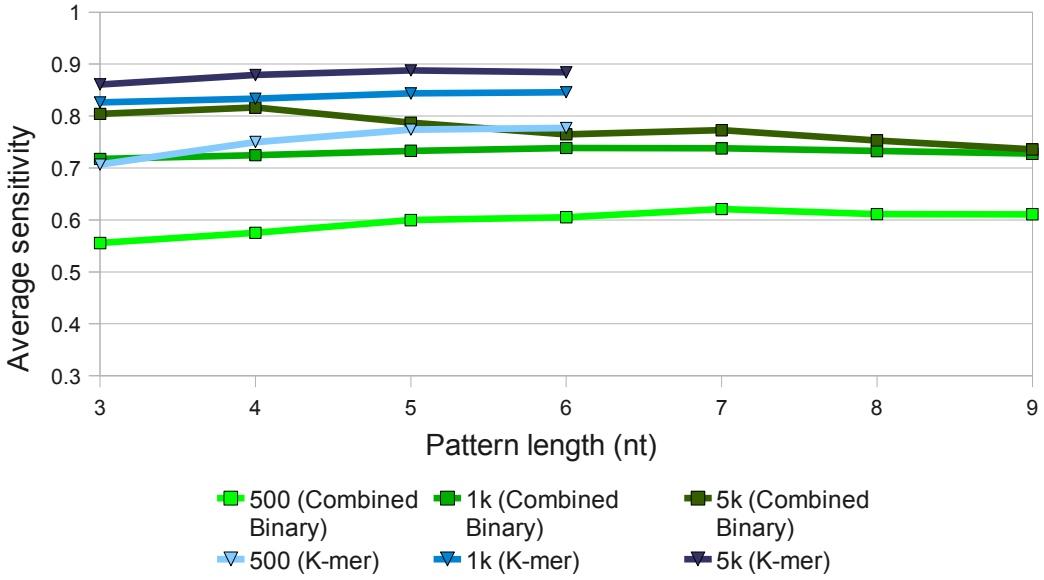
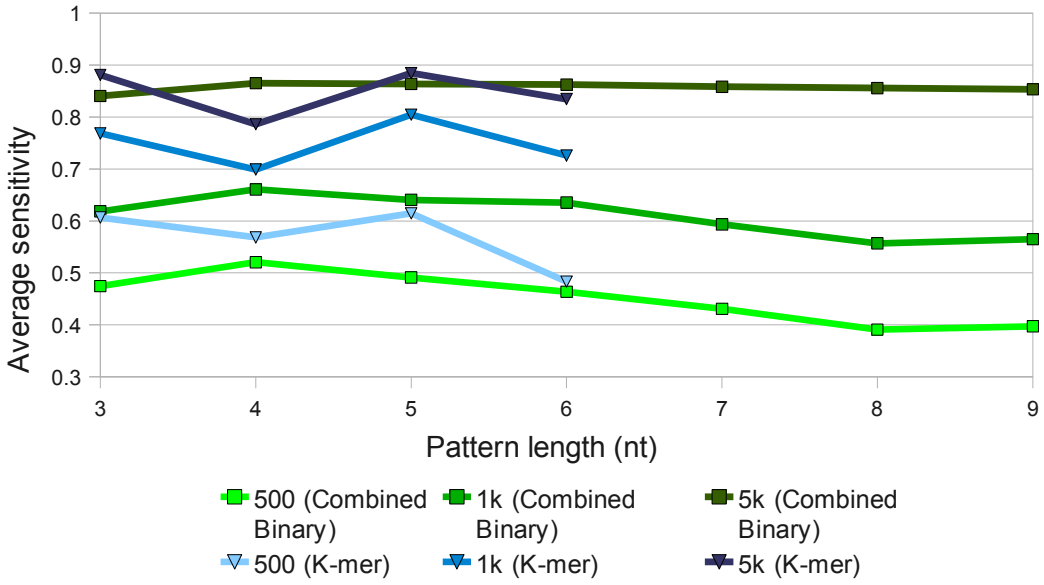


Figure 2.6: Average Sensitivity vs. Pattern Length for SVM Models Trained Using 5000 nt Fragments



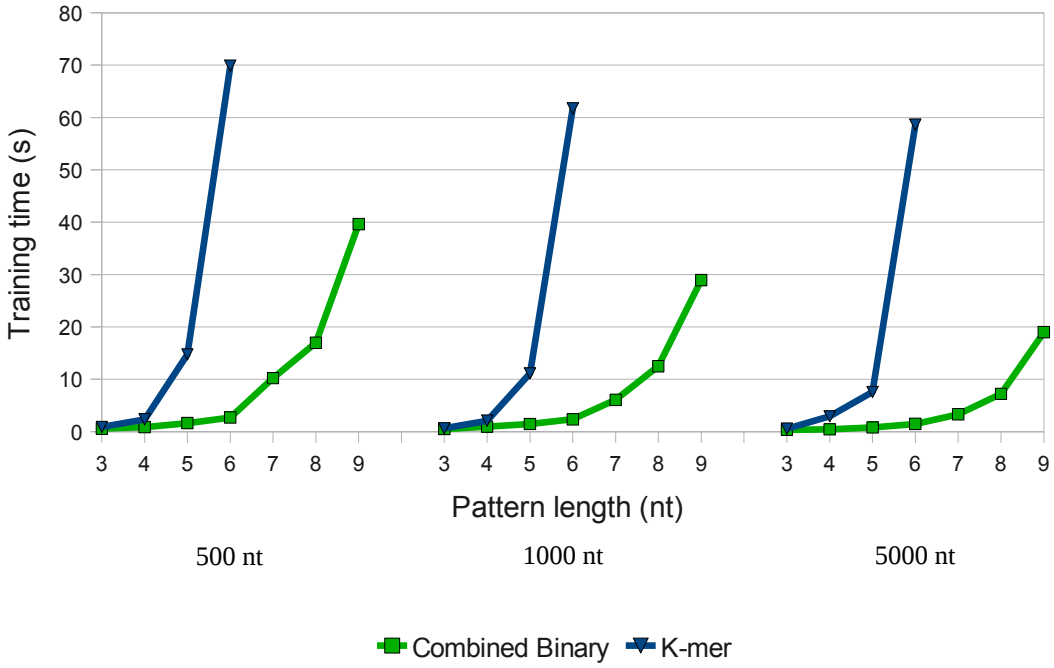
Models trained with a small fragment size are generally able to accurately classify larger fragments, in some cases classifying the larger fragments with higher sensitivities than the fragment sizes used to build the models. For example, in Figure 2.4 the k-mer classifier achieved a maximum sensitivity of 86.9% using the 5000 nt test set, despite the fact that the SVM model was trained using fragments only 500 nt in length. In contrast, models trained with a larger fragment size do not offer very high sensitivities when attempting to classify smaller fragments. Figure 2.6 demonstrates this fact quite clearly, showing that the test sets containing 5000 nt fragments gave higher sensitivities than the 500 nt and 1000 nt test sets for both the combined-binary and k-mer classifiers. In this figure, the k-mer classifier had a sensitivity of 88.5% using the 5000 nt training set, with the highest sensitivity from the other two training sets being 80.4% (1000 nt). Likewise, the combined-binary classifier showed a maximum sensitivity of 86.5% using the 5000 nt test set, but the 500 nt and 1000 nt test sets had maximum sensitivities of only 52.0% and 66.1%, respectively.

The decrease in classification sensitivity for the k-mer classifier at pattern lengths of 4 nt and 6 nt in Figure 2.6 is likely the result of an unrepresentative subset of the training data being used in the grid search, as described above for Figure 2.2.

Comparison of SVM Training and Prediction Times

Throughout all of the experiments, performance data were recorded whenever a SVM was trained or tested. Figure 2.7 illustrates the time required in order to build SVM models for the combined-binary and k-mer classifiers over a range of fragment lengths and pattern lengths. This particular set of data was obtained from the experimental trials where a fixed number of fragments (1500) were used to build the 500 nt, 1000 nt, and 5000 nt models. The combined-binary classifier had an average training time of 7.53s across all fragment and pattern lengths, with training times ranging from 0.33s – 39.64s. The k-mer classifier had comparatively higher training times, with an average of 19.44s and a range of 0.57s – 69.88s.

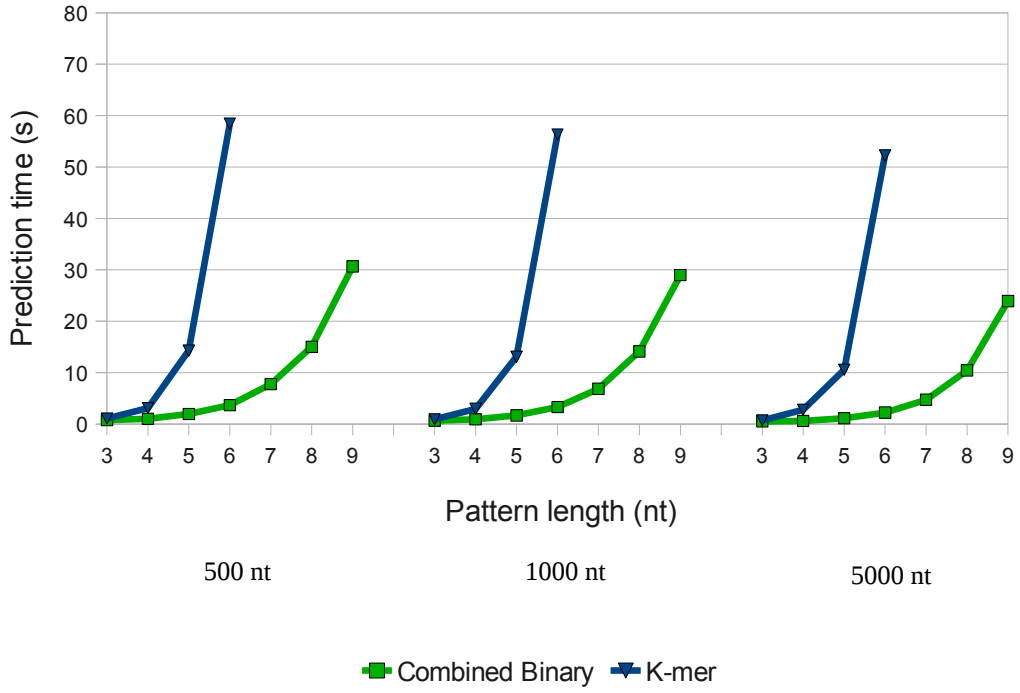
Figure 2.7: Comparison of SVM Training Times Over Varying Fragment and Pattern Lengths for the Combined-binary and K-mer Classifiers



As expected, there is a general trend of increasing training time with pattern length for both of the classifiers. Increasing the pattern length increases the number of features in the SVM training file (see Table 2.2), so it follows that libSVM would take longer to build a SVM model from a training file that contains more features. Also, Figure 2.7 shows that in all cases, more time was required to train the k-mer-based classifier for each combination of fragment length and pattern length. This may be explained by the fact that for a given pattern length, the k-mer classifier will have a much higher number of features than the alternative binary classifier, resulting in higher training times.

Perhaps somewhat counter-intuitive is the trend that the training time decreases as fragment size increases. For example, the k-mer classifier had a training time of 69.88s for the 500 nt model with a pattern length of 6, but the training time was 61.74s for the 1000 nt model using the same pattern length. Increasing the fragment length to 5000 nt further reduced the training time to 19.44s. This same trend is also observable in the SVM prediction times summarized in Figure 2.8. In both cases, it may be that the larger fragment sizes provide the SVM with a more uniform representation of the genome signatures of the fragments being classified, thus resulting in a less challenging classification task for the SVM. Additionally, the reduced training/testing times may be the result of the fact that the same number of training fragments are being used in each case, the SVMs built with the 5000 nt fragments are actually being exposed to a higher overall percentage of the given genomes being classified than the models built with smaller fragments, resulting in more accurate SVM models.

Figure 2.8: Comparison of SVM Prediction Times Over Varying Fragment and Pattern Lengths for the Combined-binary and K-mer Classifiers



Conclusions

The results presented in this chapter have demonstrated that binary-recoded DNA classifiers are in fact able to utilize genome signature in order to provide DNA sequence classification sensitivities of up to 86.5%. Unfortunately, throughout all of the experiments the k-mer classifier consistently outperformed the binary classifiers presented here, providing sensitivities of up to 89.1%. Despite the fact that binary classifiers greatly reduce the feature space, the increase in available pattern lengths facilitated by the use of binary classifiers does not offer an increase in classification sensitivity over k-mer classifiers trained with shorter patterns.

Although the binary-recoded classifiers provided similar sensitivities to the k-mer classifier at fragment sizes of 5000 nt, the classification sensitivity decreased dramatically with decreasing fragment size. This is an immense drawback for the binary classifiers, as they were intended to be applied to metagenomic data sets which often contain fragments much smaller than 5000 nt.

The results clearly demonstrate that it is advantageous to train SVM models using a short fragment length in order to ensure that the resulting models will be able to classify query fragments of various lengths. Models trained using shorter fragments have the ability to classify longer fragments without greatly sacrificing classification accuracy. Conversely, models trained using longer fragments are ineffective in attempting to classify shorter fragments.

Although DNA recoding might serve as a viable preprocessing step for other analyses, the results from this study indicate that recoding greatly reduces the SVM's ability to distinguish between genomes on the basis of genome signature. By removing compositional bias using the various simple binary recoding schemes, there is a greater chance that genomes will converge in terms of their generalized compositions. This suggests that genome signature is tightly coupled to nucleotide composition rather than pyrimidine/purine composition. Additionally, the fact that the combined-binary classifier was unable to match the k-mer classifier in terms of sensitivity indicates that k-mers

rather than individual nucleotide patterns, contribute greatly to genome signature. This would seem to suggest that codon usage biases play a dominant role in shaping a genome's nucleotide composition.

The combined-binary classifier could potentially be useful in binning genomic fragments of 5000 nt or longer. For fragments of this size, the combined-binary classifier achieved classification sensitivities comparable to those of the k-mer classifier, while offering the advantages of reduced memory utilization and running time.

Chapter 3 – SVM-mediated Pairwise Classification of 56 α-proteobacterial Genomes Based on the Tetranucleotide Profiles of Orthologous Genes

Motivation

In Chapter 2 it was shown that a simple multi-class SVM classifier is capable of distinguishing between short nucleotide sequences from 10 microbial genomes based upon their underlying k-mer frequency profiles for several values of k. Chapter 2 also demonstrated that tetranucleotide frequency profiles resulted in classification accuracies comparable to those of pentanucleotide and hexanucleotide frequency profiles. As such, tetranucleotide frequency profiles were selected for use in the current study. For multi-class classifiers, performance is often reported in terms of the average sensitivity, specificity, or balanced accuracy of a given classifier across all classes (where classes represent genomes in the present case). While these global performance measures provide convenient metrics for comparing the relative performance of different classifiers, global scores are inherently limited in that they fail to provide details about the performance of the classifier for each of the individual classes. Depending on the phylogenetic breadth of the genomes involved, the global performance scores for a given DNA classifier may be unrepresentative of the individual class-level scores. Furthermore, by considering only the global performance of a given classifier, no knowledge is gained about the specific classes that prove to be the most difficult to classify – details that might contribute to the development of a more robust classifier. For example, it is expected that the distinguishability of *Bacillus anthracis* str. Ames and *Bacillus cereus* E33L should be much lower than the distinguishability of the other genome pairs considered in Chapter 2, however the use of global accuracy scores does not provide any information about this specific comparison.

In the current study, a multi-class SVM is no longer used as the basis of the DNA classifier. Substituted in its place are a number of 2-class SVMs; a unique SVM for each of the possible pairwise groupings of the genomes used in the study. The use of 2-class SVMs allows for a much finer level of granularity when evaluating the performance of

the classifier, and avoids the shortcomings of global performance scores discussed above. An additional advantage of decomposing the classifier into multiple 2-class SVMs is that the computational effort involved in training the SVMs can be distributed across multiple CPU cores in a multi-core computer or cluster environment, an option that is not presently available when training a multi-class model with libSVM.

Rather than training each 2-class SVM using the tetranucleotide frequency profiles of genomic fragments as in Chapter 2, here we further redefine our classifier by training each 2-class SVM using only the tetranucleotide frequency profiles obtained from the putative orthologs for each pair of genomes. We focus specifically on orthologs in order to ensure that for a given genome pair, each sequence used in training the SVM to recognize one particular genome has a corresponding orthologous sequence that will be used to train the SVM to recognize the comparator genome. This strategy also attempts to avoid the confounding influence of unameliorated DNA such as viral/phage sequences which are likely to contain genome signatures quite different from the host genome, potentially decreasing the SVM's ability to distinguish between a given pair of genomes.

Sets of putative orthologs are determined using the reciprocal best hit (RBH) method with BLASTP [71; 72] as the underlying search algorithm. By determining the putatively orthologous sets of genes using RBH and BLASTP, the resulting orthologous genes may vary considerably in their nucleotide sequences while remaining significantly similar in their protein sequences due to synonymous mutations. This variation in nucleotide sequences for orthologous genes is represented in each gene's tetranucleotide frequency profile, and sufficient variation allows a SVM to distinguish between genomes on the basis of genome signature.

The reciprocal best hit method has been widely employed in order to determine putative orthologs shared between two genomes [101-104]. For a given pair of genomes $\{A, B\}$, the RBH algorithm works as follows: First, each gene in genome A is used as a query sequence against genome B using a search algorithm such as BLASTP or BLASTN [71; 72]. Subsequently, each gene in genome B is used as a query sequence in the

reciprocal search against genome *A*. A pair of genes {i_A from genome *A*, i_B from genome *B*} are deemed orthologous if i_A returns i_B as its best match when used as the query sequence against genome *B*, and likewise, i_B also returns i_A as its best match during the reciprocal query.

The goals of the present study are to model the pairwise distinguishability of genomes between 56 members of the α-proteobacteria and to identify factors that influence the level of distinguishability between a given pair of genomes. Two-class SVMs trained using the tetranucleotide frequency profiles of orthologous sequences are used to narrow the analysis to the potentially interesting and difficult-to-classify cases, and the resulting pairwise classification performances are interpreted in terms of various measures of sequence similarity.

Experimental Design

Genome Selection

A total of 56 completely sequenced α-proteobacterial genomes were selected for use in this study, representing all α-proteobacterial genomes available from NCBI as of February 27th, 2008. The class α-proteobacteria was chosen because it was known to encompass a very diverse set of species in terms of their lifestyles and environments. Many members of the class represent obligate intracellular pathogens, such as *Ehrlichia ruminantium*, *Rickettsia felis*, *Wolbachia* spp., and *Brucella* suis, which are of particular interest due to their potential for human disease or impact on agriculture. Other organisms, such as *Silicibacter TM1040* or *Rhizobium leguminosarum*, form stable endosymbiotic relationships with eukaryotic hosts. α-proteobacteria are also involved in several important metabolic processes such as photosynthesis (*Rhodobacter sphaeroides, Rhodopseudomonas palustris, Roseobacter denitrificans*) and nitrogen fixation (*Silicibacter pomeroyi*, *Rhodospirillum centenum*).

In total, the set of 56 genomes represents 44 uniquely named species within 31 distinct genera. Refer to Appendix 1 for a list of all genomes used in this study, along with relevant genomic properties.

Data Acquisition and Sequence Extraction

Protein and nucleotide sequences for all genes, as well as the taxonomic information for all genomes was acquired from NCBI as of March 1st, 2008. Genomic G+C composition for all genomes was retrieved from NCBI on March 1st, 2008. 16S rDNA sequence identity information was retrieved as a distance matrix in the DNADIST format using the MyRDP interface to the Ribosomal Database Project Release 10.1 on June 24th, 2008 [105]. In cases where a given genome contained multiple 16S rDNA genes, the first instance of a 16S rDNA sequence presented in MyRDP was selected in order to generate the 16S rDNA distance matrices.

Selection of Orthologous Genes and Calculation of Normalized BLASTP Scores

For each of the $\binom{56}{2}$ = 1540 possible 2-genome combinations of the 56 α-proteobacterial genomes, the reciprocal best hit method was used to compile sets of putatively orthologous genes. RBH queries were performed using precomputed all-vs.-all BLASTP results stored in the MOA database as of March 1st, 2008. For each pair of orthologs, the normalized-BLASTP (nBLASTP) score is defined as the average of the 2 BLASTP bitscores that contribute to the reciprocal best hit. Similarly, the average nBLASTP score for a given pair of genomes is defined as the average of all nBLASTP scores for the orthologous genes shared by the particular pair of genomes.

The total number of orthologous pairs of genes retrieved for each genome pair ranged from 442 for *Neorickettsia sennetsu* str. miyayama vs. *Zymomonas mobilis* subsp. mobilis ZM4 to 4941 for the pair of *Agrobacterium tumefaciens* str. C58 genomes. The average number of orthologous pairs across all 2-genome groupings was approximately 1129. The total amount of orthologous nucleotide sequence (counting orthologous genes from both genomes) for each of the genome pairs ranged from 892 kbp to 9.7 Mbp, with an average of 2.3 Mbp. Normalized BLASTP scores for orthologous pairs of genes ranged between 0.00695 and 1, with an average nBLASTP score of 0.459.

Ortholog Parameterization

The tetranucleotide frequency profiles (TFP) of all orthologs were calculated as follows: For a given gene *G* of length *n*, all *n*-3 overlapping windows of 4 nucleotides in width were examined in order to determine the total frequency of all 256 possible tetranucleotides {AAAA, AAAC, … TTTT} present in the gene. The overall frequencies were normalized by dividing the raw counts by the length of the given gene. The 256 resulting normalized frequencies were grouped into a vector to produce the tetranucleotide frequency profile for the gene:

$$TFP_G = \left[\frac{freq_{AAAA}}{n}, \frac{freq_{AAAC}}{n}, \frac{freq_{AAAG}}{n}, \ldots, \frac{freq_{TTTT}}{n} \right]$$

Tetranucleotide frequency profiles for each gene were calculated independently for both the coding and template DNA strands, resulting in two tetranucleotide frequency profiles for each gene. The enumeration of tetranucleotide frequencies always occurred in the 5' → 3' direction, with the first position of each tetranucleotide window oriented toward the 5' end of the gene.

Calculation of Tetramer Euclidean Distance

For a given genome pair {A,B}, the tetramer Euclidean distance (TED) was calculated as follows:

$$TED = \sqrt{(ATV_a[1] - ATV_b[1])^2 + (ATV_a[2] - ATV_b[2])^2 + \ldots + (ATV_a[256] + ATV_b[256])^2}$$

where $ATV_a[n]$ and $ATV_b[n]$ represent the n^{th} elements in the 256-element average tetranucleotide vectors (ATV) for genomes A and B, respectively. ATV for a given genome is calculated as the sum of all tetranucleotide frequency vectors for the set of orthologs in a given genome that are specific to the given genome pair, divided by the number of orthologs in the set:

$$ATV = \frac{\left[TFP_{orth0} + TFP_{orth1} + \ldots + TFP_{orthn} \right]}{n}$$

Training and Testing the SVM Models

In order to construct SVM training and testing files for a given pair of genomes, each orthologous pair of genes was first randomly assigned to one of 5 cross-validation groups. The assignment of a pair of orthologs to a given cross-validation group ensures that orthologous genes always appear together in the resulting SVM training and testing files. The tetranucleotide frequency profile for each ortholog was prepended with one of two possible class labels (0 or 1) based on the gene's source genome, in order to designate class information to the SVM during the training phase. Next, the tetranucleotide profiles for all genes assigned to a given cross-validation group were concatenated to create a set of SVM testing files, $S = \{t1, t2, t3, t4, t5\}$, where 1 through 5 identify the source cross-validation group.

For each testing file t in S, the corresponding SVM training file is formed by the concatenation of the 4 remaining SVM testing files. For instance, the training file for *t3* would consist of the concatenation of *t1*, *t2*, *t4*, and *t5*. In this manner, 5-fold leave-one-out cross-validation is easily performed by training SVM models using the 5 possible 4-element groupings of S, and then subsequently testing each model using the testing file that was excluded from the given training file.

A single grid search was performed for each pair of genomes using 500 randomly selected instances from one of the SVM testing files. The values of C and γ as determined in the grid search were used in the training of all five SVMs for the given pair of genomes. As described in Chapter 2, all SVMs in this study were built using the Radial Basis Function (RBF) kernel, as it has been shown to outperform linear kernels for tetranucleotide frequency data [38].

Training and testing of the SVMs was performed on a dual-core 3.2 Ghz desktop PC with 1.0 GB of RAM, running Ubuntu Linux version 8.04. Version 2.85 of the libSVM [100] package was used to train and test all SVMs.

For each pair of genomes, the ability of a 2-class SVM to distinguish between the genomes is defined as classification accuracy (CA). CA is calculated as the percentage of correct classifications over the 5 cross validation trials:

$$CA = \frac{C_1+C_2+C_3+C_4+C_5}{T_1+T_2+T_3+T_4+T_5} * 100$$

where C_n and T_n denote the number of correct classifications and total classifications, respectively.

Data Analysis and Selection of Outliers

Classification accuracy results for each genome pair were plotted with respect to several measures of genome similarity: 1) difference in genomic G+C content, 2) 16S rDNA sequence distance, 3) lowest common taxonomic rank, 4) average nBLASTP score, and 5) average tetramer Euclidean distance. Lowest common taxonomic rank is defined as the most specific taxonomic rank shared by both members of a given genome pair. For example, two species that share all taxonomic ranks except those of genus and species would have a corresponding LCTR of family. Several outlier genome pairs with high residuals for the CA vs. average nBLASTP model were selected for an in-depth analysis, with the ultimate goal of identifying factors that contribute to the observed increase or decrease in distinguishability relative to the model.

Results

2-class SVM models were trained for all 1540 pairwise groupings of 56 α-proteobacterial genomes. The data used to train each SVM consisted of the tetranucleotide frequency profiles of all orthologous genes shared by a given pair of genomes. Each training set was evaluated using 5-fold leave-one-out cross validation (see Experimental Design section for details) in order to determine a classification accuracy (CA) for each pair of genomes. The complete set of classification accuracies for all genome pairs was then interpreted in the context of several measures of genome similarity, and regression analysis was used to fit models, when possible. The majority of genome pairs are easily distinguished by the SVMs, providing a mean CA of 97.2% across all comparisons. The total range of CA values is 49.84% to 100%.

Regression analysis was performed in order to fit a quadratic model to the

classification accuracy vs. average nBLASTP data set, and a logarithmic model to the classification accuracy vs. 16S rDNA distance data set. The average nBLASTP model gave an R^2 value of 0.7761 and p-value < 2.2e-16, whereas the 16S rDNA distance model gave an R^2 of 0.7132 and a p-value < 2.2e-16. Given that only orthologs were used in this study, one would expect that the average nBLASTP scores would provide the most accurate model, as suggested by the differences in R^2 values. For this reason, residuals for the average nBLASTP model were used to select a set of outliers that were easier or more difficult to classify than suggested by the model. Characteristics of these outlier genomes were subsequently investigated in order to try to determine specific factors that contribute to the residual classification accuracy. Regression analyses were performed using the R statistical computing package [106] version 2.8.

The relationship between CA and the average nBLASTP scores for the set of orthologs shared by each pair of genomes is depicted in Figure 3.1. For genome pairs with an nBLASTP score less than 0.7, the SVMs are always able to distinguish between the genomes with greater than 80% accuracy, and for nBLASTP scores less than 0.45, classification accuracy always exceeds 87.8%. Conversely, the average CA for genome pairs with nBLASTP scores above 0.7 is 67.3%, with no pairs ever exceeding 90.4%. The best-fit quadratic model ($R^2 = 0.7761$) is shown as a solid grey line in Figure 3.1. The model is useful in helping to identity outlier genome pairs that are easier or more difficult to classify than would be expected given their average nBLASTP scores. Several outlier pairs (denoted by red symbols) were selected in order to try to identify genomic characteristics that may influence genome distinguishability. Refer to Table 3.1 for a list of all outlier pairs and their associated residuals.

CA is directly proportional to 16S rDNA distance, with an increase in 16S rDNA distance leading to a corresponding increase in CA (Figure 3.2). Genomes with less than 5% difference in their 16S rDNA genes are in general more difficult to classify, with an average CA of 66.1% for the 79 pairs in this category. Genomes with less than 1% difference in their 16S genes are essentially indistinguishable by the SVMs, with an average CA of only 54.77% for these 15 pairs. Above a 16S rDNA distance of about 5%,

all genome pairs are classified with high accuracy, giving a mean CA of 98.3% and no pairs falling below 85.57%. A best-fit logarithmic model gave an R^2 of 0.7132, slightly less than the model provided by the average nBLASTP scores above. The difference in R^2

Figure 3.1: Classification Accuracy Versus Average nBLASTP for all Genome Pairs.

Regression analysis was used to fit a quadratic model (R^2 = 0.7761, p-value < 2.2e-16), represented by the solid grey line. Red symbols are used to denote selected outliers, as follows: **crosses**: *Anaplasma phagocytophilum* vs. *Neorickettsia sennetsu*, **triangle**: *Silicibacter pomeroyi* vs. *Silicibacter* sp. TM1040, **squares**: *Ehrlichia canis* str. Jake vs. *E. ruminantium* str. Welgevonden v1, *E. canis* str. Jake vs. *E. ruminantium* str. Welgevonden v2, *E. canis* str. Jake vs. *E. ruminantium* str. Gardel, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v1, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v2, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Gardel, **circles**: *Rickettsia prowazekii* str. Madrid E vs. *R. felis* URRWXCal2, *R. conorii* str. Malish 7 vs. *R. prowazekii* str. Madrid E, *R. conorii* str. Malish 7 vs. *R. typhi* str. Wilmington, *R. typhi* str. Wilmington vs. *R. felis* URRWXCal2.

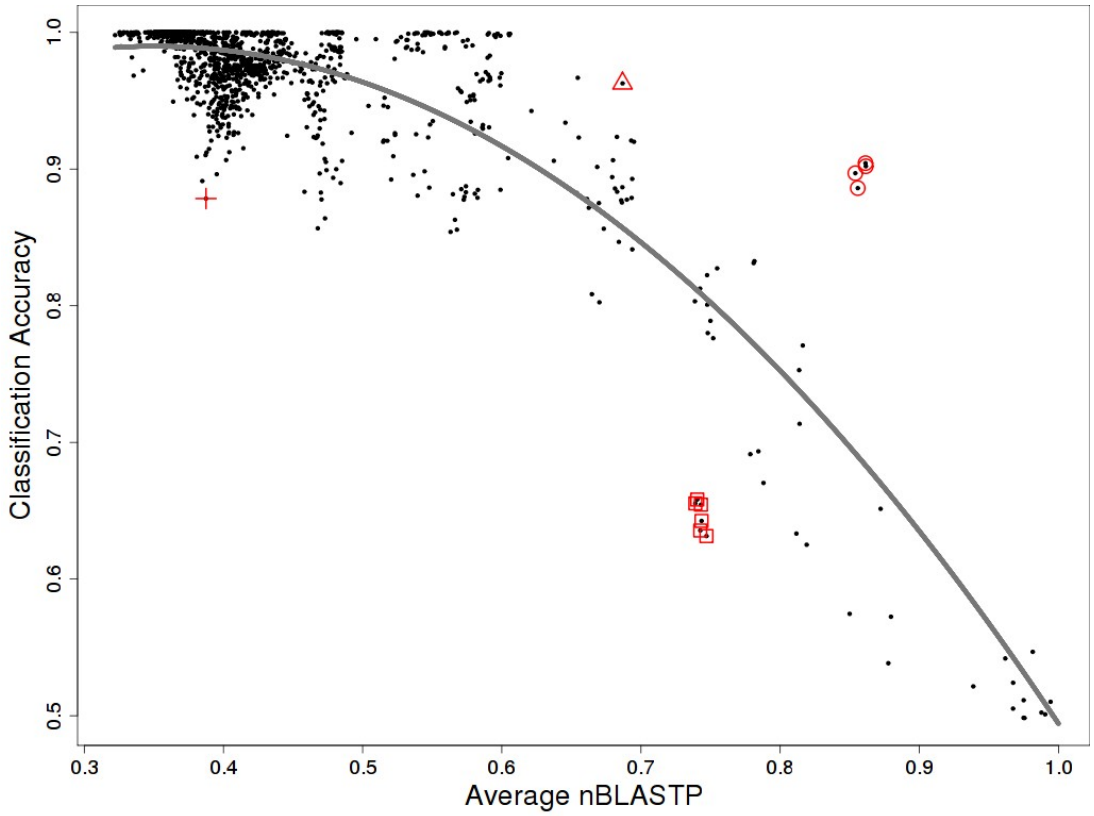


Figure 3.2: Classification Accuracy Versus 16S rDNA Distance

Classification accuracy for each pair of genomes was plotted with respect to 16S rDNA sequence distance as determined from the uncorrected distance matrix retrieved from RDP. Red symbols are used to denote outliers selected from the CA versus NBLASTP model, as follows: **crosses**: *Anaplasma phagocytophilum* vs. *Neorickettsia sennetsu*, **triangle**: *Silicibacter pomeroyi* vs. *Silicibacter* sp. TM1040, **squares**: *Ehrlichia canis* str. Jake vs. *E. ruminantium* str. Welgevonden v1, *E. canis* str. Jake vs. *E. ruminantium* str. Welgevonden v2, *E. canis* str. Jake vs. *E. ruminantium* str. Gardel, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v1, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v2, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Gardel, **circles**: *Rickettsia prowazekii* str. Madrid E vs. *R. felis* URRWXCal2, *R. conorii* str. Malish 7 vs. *R. prowazekii* str. Madrid E, *R. conorii* str. Malish 7 vs. *R. typhi* str. Wilmington, *R. typhi* str. Wilmington vs. *R. felis* URRWXCal2.

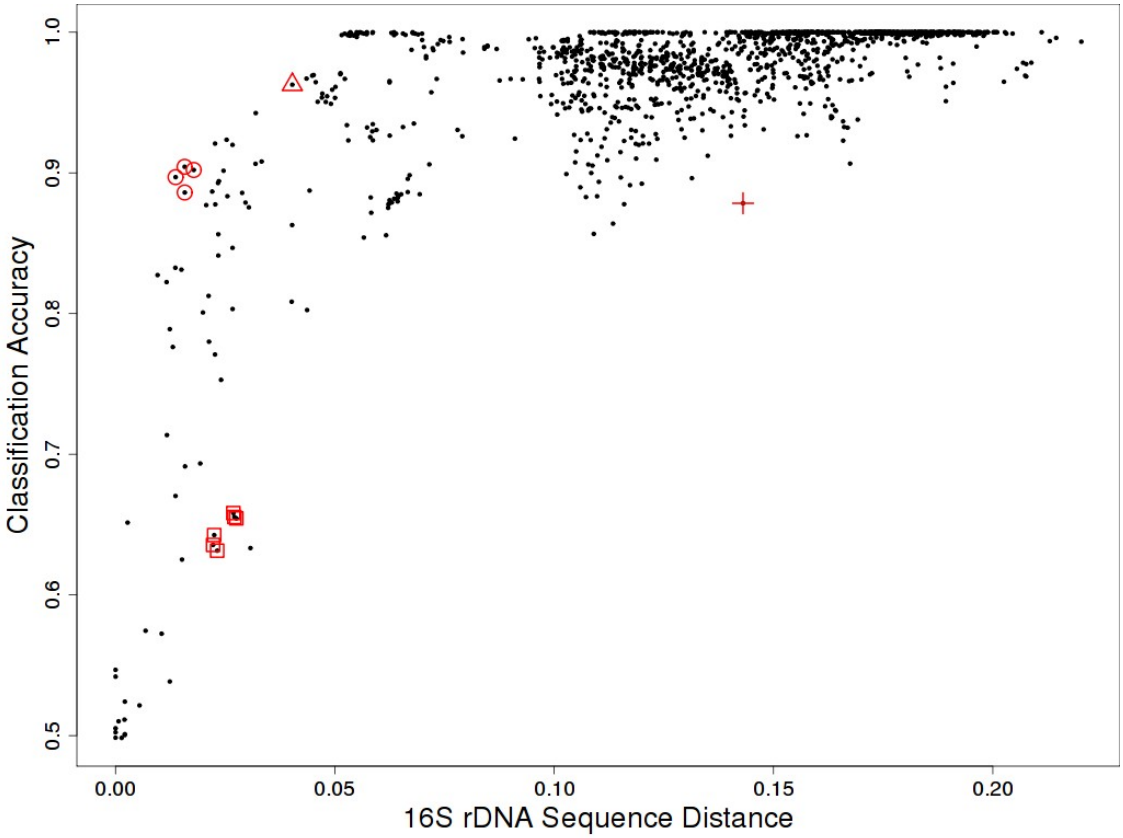


Table 3.1: Outlier Pairs Selected for Further Investigation

Residual CA refers to the residual CA as determined by the CA vs. average nBLASTP model.

Genome Pair	CA	Residual CA	# Orthologs	Average nBLASTP	16S rDNA Distance	G+C Distance	Tetramer Distance
Anaplasma phagocytophilum vs. *Neorickettsia sennetsu*	0.878	-0.110	572	0.387	0.143	0.005	0.0140
Silicibacter pomeroyi vs. *Silicibacter* sp. TM1040	0.963	**0.106**	2667	0.687	0.040	0.040	0.0206
Ehrlichia canis str. Jake vs. *E. ruminantium* str. Welgevonden v1	0.654	-0.154	790	0.743	0.028	0.015	0.0044
E. canis str. Jake vs. *E. ruminantium* str. Welgevonden v2	0.655	-0.157	792	0.739	0.027	0.015	0.0046
E. canis str. Jake vs. *E. ruminantium* str. Gardel	0.658	-0.153	793	0.740	0.027	0.015	0.0046
E. chaffeensis str. Arkansas vs. *E. ruminantium* str. Welgevonden v1	0.631	-0.174	825	0.747	0.023	0.026	0.0041
E. chaffeensis str. Arkansas vs. *E. ruminantium* str. Welgevonden v2	0.636	-0.174	793	0.743	0.022	0.026	0.0042
E. chaffeensis str. Arkansas vs. *E. ruminantium* str. Gardel	0.643	-0.166	796	0.743	0.022	0.026	0.0042
Rickettsia prowazekii str. Madrid E vs. *R. felis* URRWXCal2	0.902	**0.219**	802	0.862	0.018	0.035	0.0107
R. conorii str. Malish 7 vs. *R. prowazekii* str. Madrid E	0.886	**0.196**	790	0.856	0.016	0.034	0.0100
R. conorii str. Malish 7 vs. *R. typhi* str. Wilmington	0.897	**0.205**	791	0.854	0.014	0.035	0.0103
R. typhi str. Wilmington vs. *R. felis* URRWXCal2	0.904	**0.221**	805	0.861	0.016	0.036	0.0110
Average across all genome pairs	0.972	1.29E-020	1129	0.424	0.138	0.156	0.0507

values is likely due to the fact that the average nBLASTP scores are tightly coupled to the putative orthologs that are being classified by the SVM, because it is the underlying BLASTP scores that are initially used in the RBH queries to define the set of orthologs. 16S rDNA distance, on the other hand, represents a measure of genome similarity based on a single highly conserved gene. Despite the relative simplicity of determining 16S rDNA distance in comparison to average nBLASTP scores, and the portions of each genome pair that have been excluded due to their non-orthologous nature, 16S rDNA distance is still a reasonable predictor of CA for this data set.

Genomic G+C distance appears to define a minumum bound on CA (Figure 3.3). Unlike Figures 3.1 and 3.2, where 95% or better CA is only achievable within a small range of nBLASTP scores or 16S rDNA distances, pairs of genomes with equivalently high CA values are found throughout the entire range of G+C distances. For example, the 14 genome pairs with identical G+C content have a CA range of 49.84% - 97.75%, with a mean CA of 68.02%. Genome pairs with a G+C difference above 10% (774 pairs in total) range in CA from 98.2% - 100%, with a mean of 99.88%.

Classification accuracy is compared with the tetramer Euclidean distance for each genome pair in Figure 3.4. Across all genome pairs, the mean tetramer distance is 0.0507 with a range of 0.0001 to 0.1107. Unlike genomic G+C content which appears to impose only a lower bound on CA, tetramer distance appears to impose both upper and lower bounds on CA. For genome pairs with negligible differences in average tetramer composition of their shared orthologs, CA is approximately 50%. As tetramer distance increases from 0 to 0.015, CA increases approximately linearly from 0% - 93.91%. Tetramer distance values in the range of 0.015 – 0.04 show moderate variability in CA, with CA values ranging from 87.71% - 100% (mean: 96.88%). Beyond a tetramer distance of 0.04, the mean CA is 99.9%, with CA never falling below 97.9%.

CA can be interpreted in terms of the taxonomic relatedness of each pair of genomes (Figure 3.5). When the CA vs. nBLASTP results are partitioned by the lowest common taxonomic rank (LCTR) of each genome pair, there is a trend of decreasing CA as LCTR becomes more specific. For the 1141 genome pairs with a LCTR of 'Class', CA

Figure 3.3: Classification Accuracy Versus Genomic G+C Distance

CA was plotted against genomic G+C distance for all pairs of genomes. Red symbols are used to denote outliers selected from the CA versus NBLASTP model, as follows: **crosses**: *Anaplasma phagocytophilum* vs. *Neorickettsia sennetsu*, **triangle**: *Silicibacter pomeroyi* vs. *Silicibacter* sp. TM1040, **squares**: *Ehrlichia canis* str. Jake vs. *E. ruminantium* str. Welgevonden v1, *E. canis* str. Jake vs. *E. ruminantium* str. Welgevonden v2, *E. canis* str. Jake vs. *E. ruminantium* str. Gardel, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v1, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v2, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Gardel, **circles**: *Rickettsia prowazekii* str. Madrid E vs. *R. felis* URRWXCal2, *R. conorii* str. Malish 7 vs. *R. prowazekii* str. Madrid E, *R. conorii* str. Malish 7 vs. *R. typhi* str. Wilmington, *R. typhi* str. Wilmington vs. *R. felis* URRWXCal2.

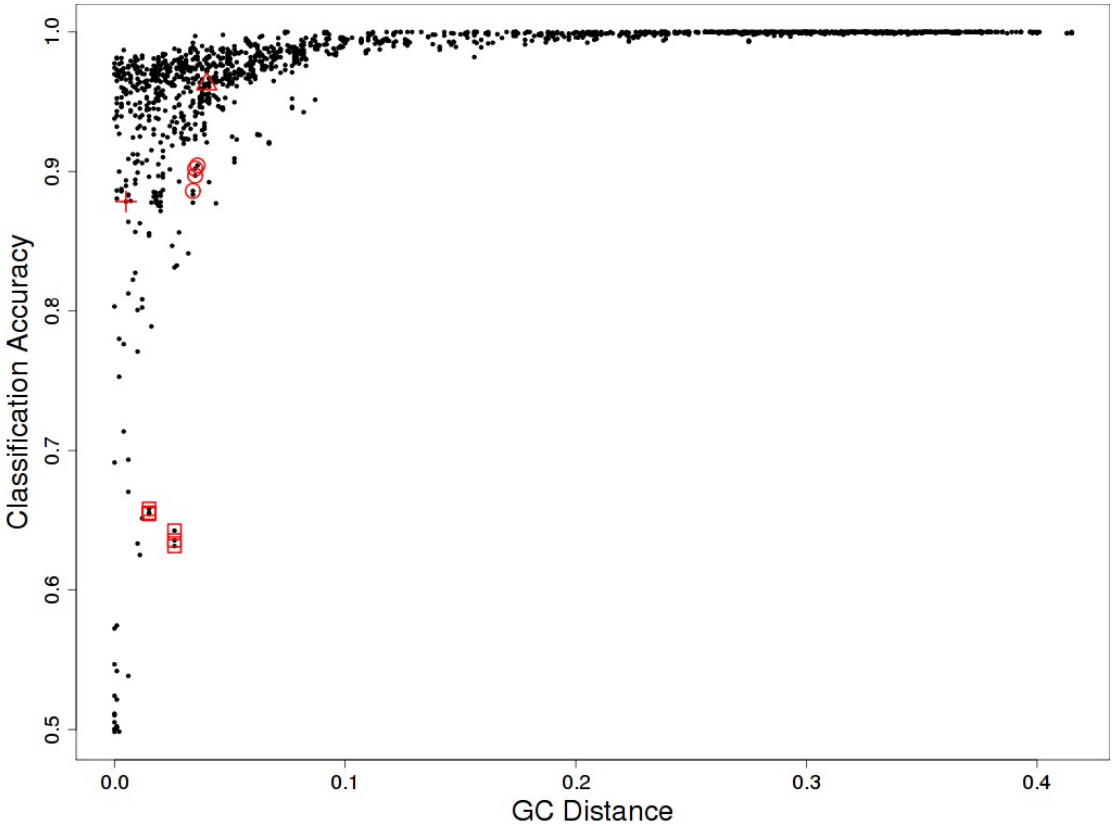


Figure 3.4: Classification Accuracy Versus Average Tetramer Distance

Tetramer distance was calculated using the euclidean distance bewteen the average tetranucleotide profile for the set of orthologs for each genome in a given genome pair. Red symbols are used to denote outliers selected from the CA versus NBLASTP model, as follows: **crosses**: *Anaplasma phagocytophilum* vs. *Neorickettsia sennetsu*, **triangle**: *Silicibacter pomeroyi* vs. *Silicibacter* sp. TM1040, **squares**: *Ehrlichia canis* str. Jake vs. *E. ruminantium* str. Welgevonden v1, *E. canis* str. Jake vs. *E. ruminantium* str. Welgevonden v2, *E. canis* str. Jake vs. *E. ruminantium* str. Gardel, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v1, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v2, *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Gardel, **circles**: *Rickettsia prowazekii* str. Madrid E vs. *R. felis* URRWXCal2, *R. conorii* str. Malish 7 vs. *R. prowazekii* str. Madrid E, *R. conorii* str. Malish 7 vs. *R. typhi* str. Wilmington, *R. typhi* str. Wilmington vs. *R. felis* URRWXCal2.

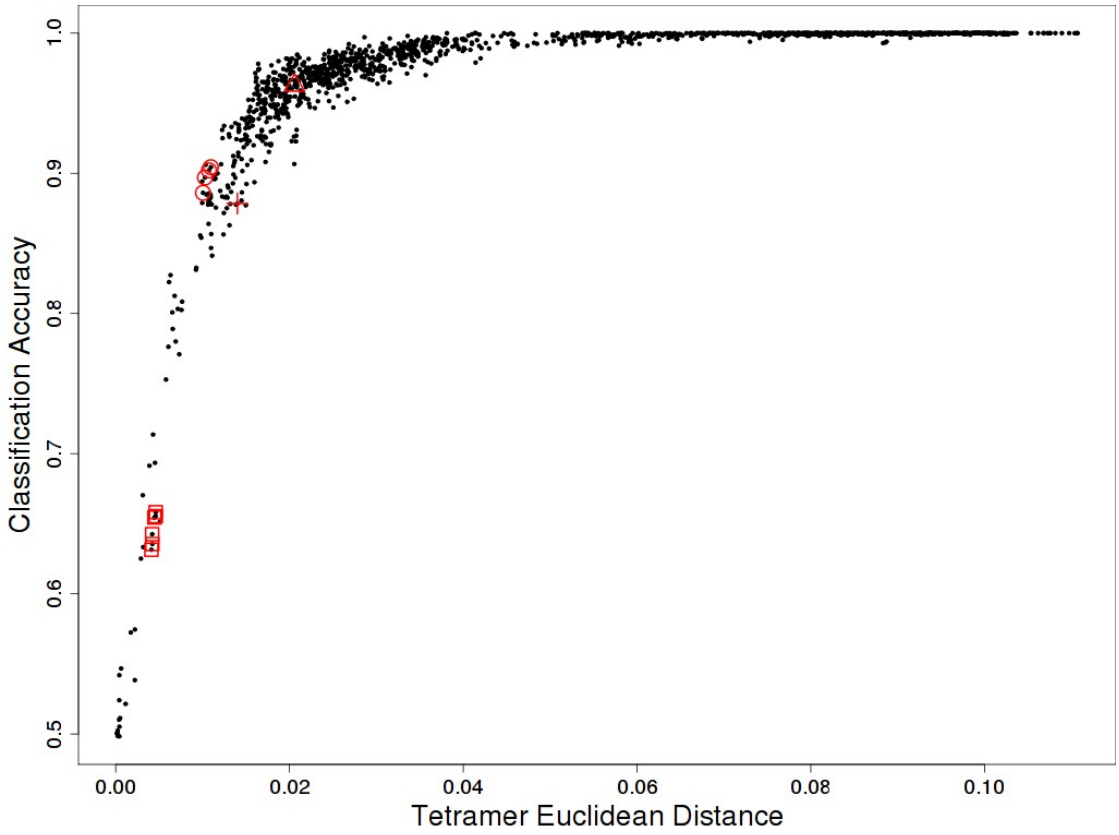
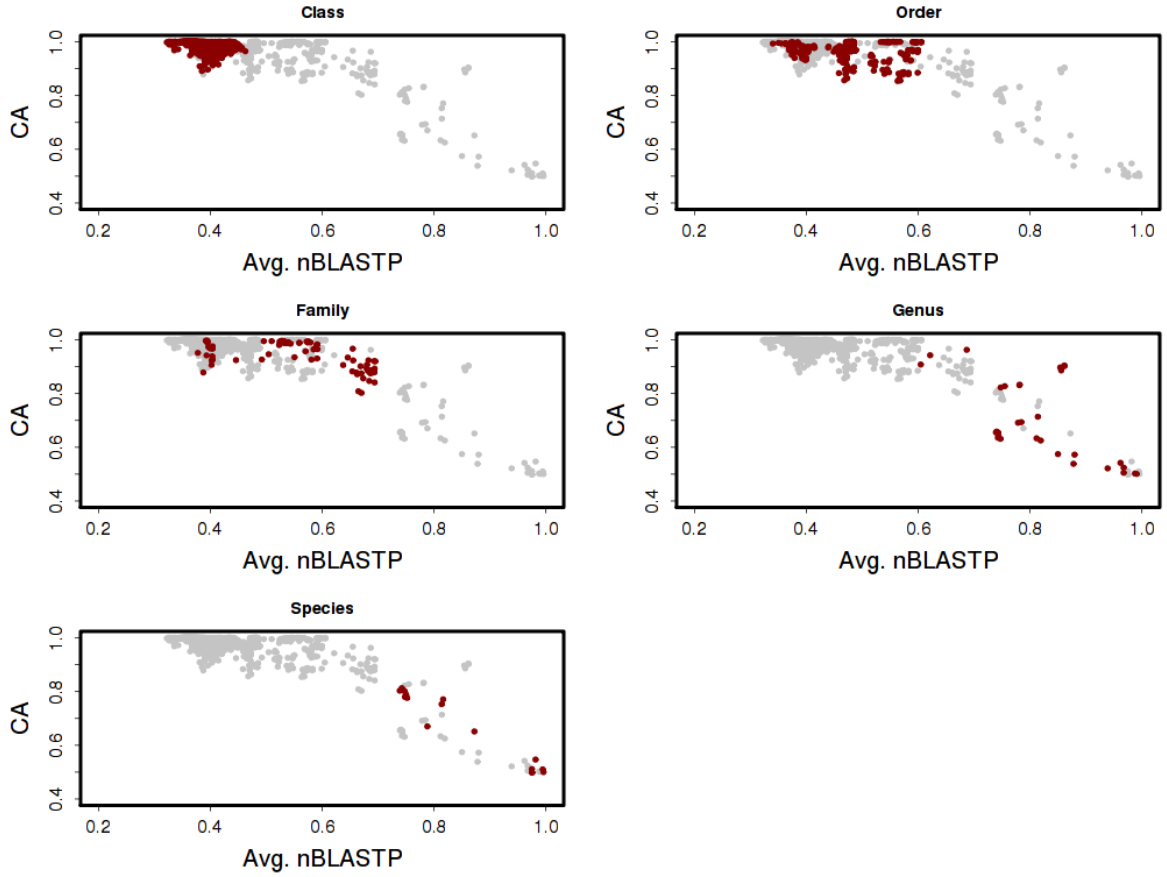


Figure 3.5: Classification Accuracy Versus Average nBLASTP, Partitioned by Lowest Common Taxonomic Rank

Classification accuracy versus average nBLASTP results were partitioned based upon the most specific taxonomic rank shared by both members of each genome pair. Red dots indicate the results that are specific to the given taxonomic rank.

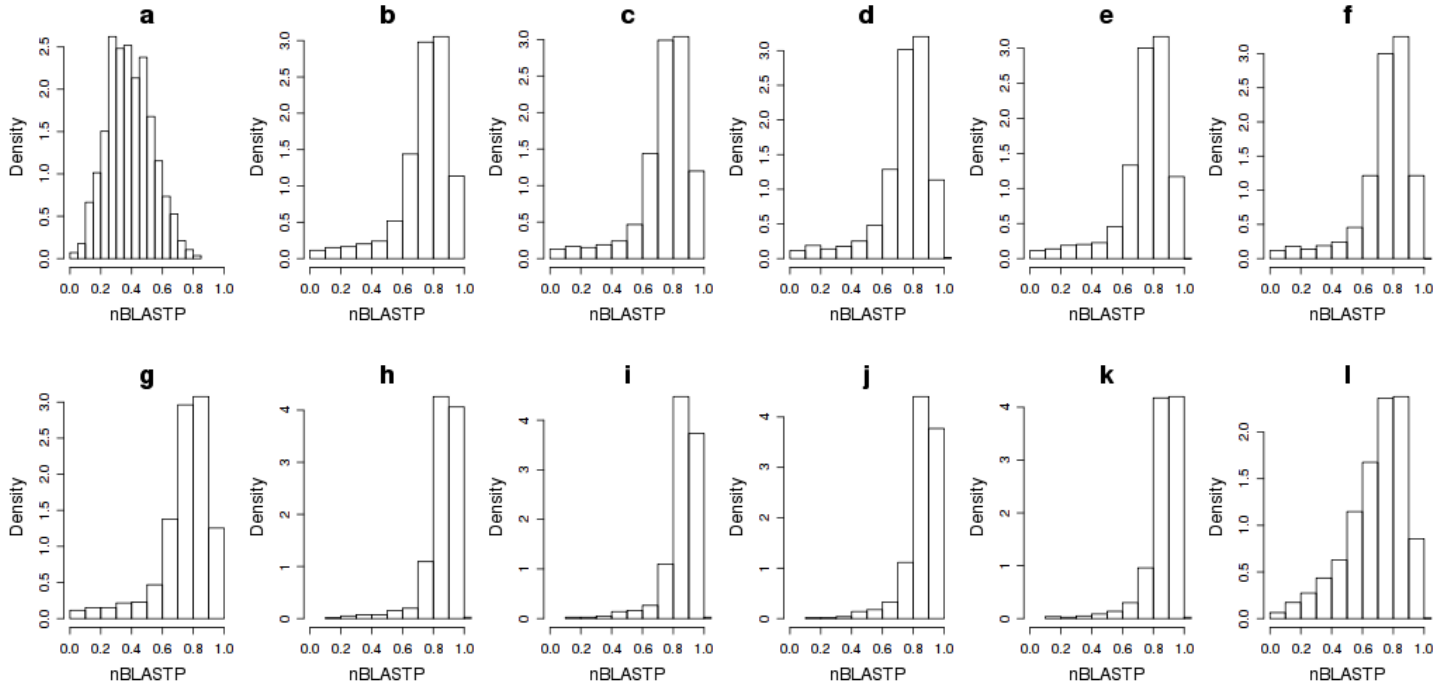


ranges from 89.13% - 100% with a mean CA of 98.81%. At a LCTR of 'Order', the 275 corresponding genome pairs have a mean CA of 96.24%, with values ranging from 77.44% - 100%. Similarly, the pairs with a LCTR of 'Family' range in CA from 80.25% - 99.66%, with a mean CA of 94%. CA drops considerably for genome pairs at the LCTR of 'Genus' and 'Species'. For 'Genus', the 35 corresponding pairs range in CA from 50.1% - 96.27%, with a mean CA of 70.10%. The 16 conspecific genome pairs range in CA from 49.84% - 81.26%, with a mean of 66.71%. Among the conspecific genome pairs, the least distinguishable are strains of *Rhodobacter sphaeroides* (CA = 49.85%), *Ehrlichia ruminantium* (CA = 49.84%), and *Brucella abortus* (CA = 50.03%). The most distinguishable genome pairs at the LCTR of 'Species' include 10 pairs of strains of *Rhodopseudomonas palustris* (CA range: 65.14% - 81.26%). The next most distinguishable conspecific genome pair contains two strains of *Agrobacterium tumefaciens*, which are distinguishable at a CA of 54.67%.

The distribution of nBLASTP scores for a given pair of genomes can shed light on the overall similarity between orthologous genes in a pair of genomes. Figure 3.6 illustrates the population density distributions of the nBLASTP scores for the sets of orthologous genes shared by each of the outlier genome pairs listed in Table 3.1. For closely related genomes (panels b-l), the nBLASTP scores assume a negatively skewed unimodal distribution with a peak centered at a nBLASTP value of 0.8 – 0.9, indicating that such genome pairs have a higher proportion of orthologs that are similar in protein sequence. The lone pair of genomes involving two distinct genera, *A. phagocytophilum* vs. *N. sennetsu* (panel a) has a normal distribution centered at a nBLASTP score of 0.4, suggesting that the orthologs shared by these genomes differ considerably in terms of their protein sequences.

Figure 3.6: Distribution of RBH nBLASTP Scores for each Genome Pair

Panels a-m show the population density distribution of nBLASTP scores for the set of orthologs shared by each genome pair. **a**: *Anaplasma phagocytophilum* vs. *Neorickettsia sennetsu*, **b**: *E. canis* str. Jake vs. *E. ruminantium* str. Welgevonden v2, **c**: *E. canis* str. Jake vs. *E. ruminantium* str. Gardel, **d**: *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Welgevonden v1, **e**: *E. chaffeensis* str. Arkansas vs. *E. ruminantium* str. Gardel, **f**: Arkansas vs. *E. ruminantium* str. Welgevonden v1, **g**: *E. canis* str. Jake vs. *E. ruminantium* str. Welgevonden v1, **h**: *R. typhi* str. Wilmington vs. *R. felis* URRWXCal2, **i**: *R. conorii* str. Malish 7 vs. *R. typhi* str. Wilmington, **j**: *R. conorii* str. Malish 7 vs. *R. prowazekii* str. Madrid E, **k**: *Rickettsia prowazekii* str. Madrid E vs. *R. felis* URRWXCal2, **l**: *Silicibacter pomeroyi* vs. *Silicibacter* sp. TM1040.



Outlier Comparison

Based on the residual CA values from the CA vs. nBLASTP model, a set of 12 outlier genome pairs was selected for further investigation (see Table 3.1). Outliers are of particular interest because they may draw attention to features that either improve (positive outliers) or confound (negative outliers) classification. A total of five positive outliers are included in the set, consisting of a single comparison between *Silicibacter pomeroyi* and *Silicibacter* sp. TM1040, and 4 comparisons between various species of rickettsia (*R. prowazekii* vs. *R. felis*, *R. prowazekii* vs. *R. conorii*, *R. conorii* vs. *R. typhi*, and *R. typhi* vs. *R. felis*). These 4 *Rickettsia* outliers form a coherent cluster that is visible in Figures 3.1-4. Negative outliers consist of a comparison between *Anaplasma phagocytophilum* and *Neorickettsia senettsu*, and 6 comparisons between *Ehrlichia* species (3 strains of *E. ruminatium* compared with *E. canis* and *E. chaffeensis*, respectively). Similar to the *Rickettsia* outliers, the *Ehrlichia* comparisons are also clustered in Figures 3.1-4.

Many of the genomes involved in the outlier comparisons share several broad-level genomic characteristics. With the exception of the two *Silicibacter* species, all remaining genomes belong to the order Rickettsiales, and represent obligate intracellular pathogens of mammalian hosts. These pathogens are typically spread via arthropod vectors (ticks, fleas) and primarily infect macrophages, neutrophils, or endothelial cells, where they either live freely within the cytosol or take refuge within vacuoles. Inside an infected cell, these organisms rely on type IV secretion systems in order to exchange DNA and other substrates with the host cell. [66; 107-113]

In general, the genomes of intracellular pathogens are greatly reduced in size relative to the genomes of free-living bacteria, and tend to have relatively low G+C content [114]. The reduced nature of these genomes has also resulted in fewer tRNA genes than their free-living counterparts, and many of the genomes have lost genes for entire pathways relating to nucleotide and amino acid biosynthesis, resulting in an obligate reliance on the host cell to supply these materials [113; 115]. In many intracellular pathogens, genes that are normally found grouped within operons in free-

living bacteria are found scattered throughout the genome, indicating that such genomes have an increased likelihood of undergoing rearrangement [109; 113]. Population bottlenecks experienced by obligate intracellular pathogens may lead to rapid gene loss and fixation of mutations that are uncharacteristic of populations of free-living bacteria [108; 114; 116]. It is also believed that limited exposure to other bacteria provides intracellular pathogens with less opportunity for the exchange of genetic material via LGT [115], although a small number of LGT events have been identified in *Rickettsia* massiliae. [117].

All of these features could in one way or another contribute to the residual CA observed for each of the selected outlier pairs. The results for each of the groups of outliers are considered in terms of their specific genomic characteristics in the following section.

Anaplasma phagocytophilum* vs. *Neorickettsia sennetsu

The comparison of *A. phagocytophilum* and *N. sennetsu* represents an interesting negative outlier. Figure 3.6a shows the distribution of nBLASTP scores for this pair of genomes, highlighting its relatively low nBLASTP scores compared to all of the other outliers. With a low average nBLASTP score of 0.387, the corresponding CA of 87.8% is significantly less than that of other genome pairs with comparable nBLASTP scores (see Figure 3.1). For instance, the 20 other genome pairs with nBLASTP scores between 0.386 and 0.388 have a mean CA of 97.67%, with the next lowest CA being 91.02%. Likewise, the CA vs. 16S rDNA distance plot (Figure 3.2) also suggests that this outlier should be expected to have a higher CA; the 20 genome pairs with comparable rDNA distances in the range of 0.142 – 0.144 show a mean CA of 98.11%. When examined in terms of genomic G+C content, Figure 3.3 shows that *A. phagocytophilum* is very close in G+C composition to *N. sennetsu* (G+C distance = 0.005), suggesting that convergence of G+C composition may play some role in reducing the distinguishability of these genomes. It should be noted, however, that genome pairs with G+C distances < 0.005 were able to achieve CA values as high as 98.72% in some cases, so the effect of G+C convergence on distinguishability may be minimal in this instance.

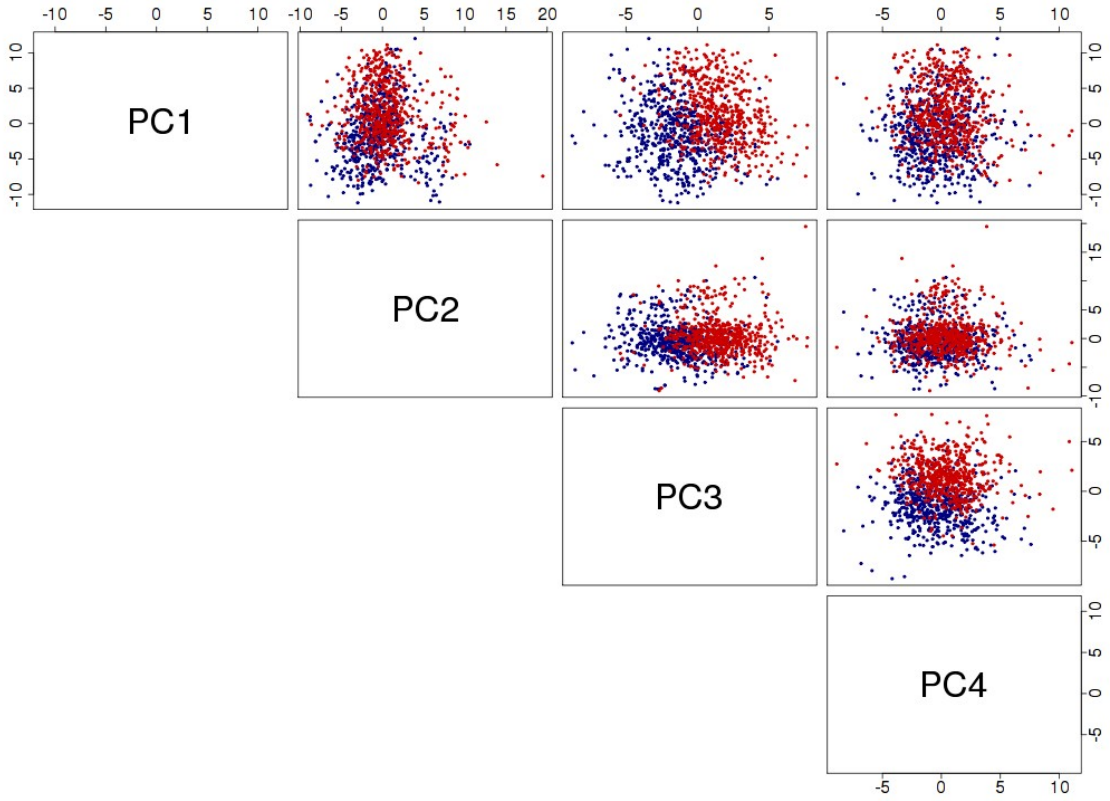
In Figure 3.4, it is apparent that this outlier falls into the lower 25% of tetramer distances, despite its below-average nBLASTP scores. This suggests that although the orthologs shared by this genome pair have diverged in terms of their protein sequences since the most recent common ancestor, the tetranucleotide composition has not diverged substantially. Principal component analysis has previously been used to distinguish between genomes on the basis of oligonucleotide frequency data [65]. In Figure 3.7, the first 4 principal components of the tetranucleotide frequency data for this genome pair are compared using pairwise scatterplots. Although the data are somewhat separable on the basis of the principal components, Figure 3.7 reiterates the fact that there is substantial overlap in the tetranucleotide frequency profiles of these two genomes.

There are several possible explanations for the lack of tetramer divergence (or conversely, increase in tetramer convergence) observed for this outlier pair. Both of these genomes are from intracellular pathogens that reside in vacuoles within the host cell [108], and thus may have very limited opportunity to acquire new genome sequence via LGT. It is also possible that similarities in niche have influenced the convergence of the tetranucleotide profiles of these genomes. For example, Willenbrock et al demonstrated that codon usage bias provides sufficient signal to cluster 323 microbial genomes into groups based on the lifestyle of the organisms [44]. In a separate study, environment was shown to have a significant influence on G+C content and amino acid composition [48].

Additionally, both genomes have very few mobile elements (no intact prophage or transposable elements), a feature which may serve to reduce the variability in tetramer composition [108]. Similarly, a lack of several DNA repair enzymes in *N. sennetsu* may also contribute to reduced tetramer divergence, as a lack of DNA repair mechanisms has previously been shown to have a direct influence on genome composition [118].

Figure 3.7: Scatterplots for the First Four Principal Components of the Tetranucleotide Frequency Profiles for A. phagocytophilum vs. N. Sennetsu

The 6 panels in this figure represent the pairwise scatterplots of the first four principal components of the tetranucleotide frequency profiles for the orthologs shared by *A. phagocytophilum* (blue dots) and *N. sennetsu* (red dots).



Silicibacter pomeroyi* vs. *Silicibacter sp. TM1040

The silicibacters are an interesting positive outlier, representing the sole comparison of free-living bacteria included among the set of outliers. Unlike the other outliers, the silicibacters have larger genomes (> 4 Mbp), share considerably more orthologous genes (approximately 3X the number of orthologs as the other outliers), and have a much higher G+C content (> 60%). Although they share a moderately high average nBLASTP score, this pair of genomes also has a remarkably high CA of 96.3%. Figure 3.1 shows that this outlier falls at the extreme upper range of CA among other points with comparable nBLASTP scores. Likewise, this pair also demonstrates higher-than-expected CA (see Figure 3.2 when compared against genome pairs with similar 16S rDNA distances. The mean CA for genome pairs with comparable 16S rDNA distances (0.035 – 0.045) was 88.29%, with a range of 80.25% - 96.91%.

In terms of composition, the silicibacters exhibit both the highest G+C distance (G+C=0.040, Figure 3.3) and highest tetramer distance (tetramer distance=0.0206, Figure 3.4) of all the outliers. These compositional measures suggest that although the shared orthologs do not differ considerably in protein sequence as indicated by the high nBLASTP scores (Figure 3.1, Figure 3.6l), the tetranucleotide compositions of the underlying nucleotide sequences have diverged to a much greater extent. One possible explanation is that one or both of the genomes have accumulated an abundance of synonymous mutations, such that the nucleotide sequences have diverged while retaining the integrity of the protein sequences. Different evolutionary strategies may also have influenced the divergence in tetranucleotide usage. *Silicibacter* sp. TM1040 is a free-living organism, but is able to form an endosymbiotic relationship with dinoflagellates, resulting in a biofilm on the surface of the host cells [119]. In constrast, *S. pomeroyi* is not a facultative endosymbiont, but has instead adopted a lithoheterotrophic existence, acquiring genes for numerous metabolic pathways that provide several alternative energy sources depending on nutrient availability [120]. This sharp contrast in evolutionary strategies may have helped to pull the tetranucleotide compositions of these organisms in different directions, contributing to the positive residual CA. In support of this notion, a

previous comparison of the genomes of two strains of *Prochlorococcus* suggested that niche differentiation has greatly contributed to genome divergence in these organisms [121].

Ehrlichia spp.

In Figure 3.1, it is apparent that the 6 *Ehrlichia* outliers all have CA values well below what might be expected for their corresponding average nBLASTP scores. For instance, the outliers have CA values in the range of 63.1% - 65.8%, whereas the other 6 comparisons with comparable nBLASTP scores (0.73 – 0.75) range in CA from 78% - 82.2%. Despite having below average nBLASTP scores which suggest divergence in the protein sequences of the orthologous genes shared by each pair of outliers, the outliers all have below average 16S rDNA distances (Figure 3.2), G+C distances (Figure 3.3), and tetramer Euclidean distances (Figure 3.4) indicating that the nucleotide sequences of the orthologous genes have not diverged to a great extent.

Since each of the average nBLASTP scores used in Figure 3.1 is calculated from the individual nBLASTP scores for each set of orthologs, an unusual distribution of the individual nBLASTP scores might mean that the average value is unrepresentative for a given pair of genomes. In Figure 3.6, panels b-g show that the nBLASTP scores for all of the Ehrlichia outliers have skewed normal distributions with peaks centred around 0.7, thus removing the possibility that the average nBLASTP scores are grossly unrepresentative for these pairs of genomes.

As with *A. phagocytophilum* and *N. sennetsu*, all of the *Ehrlichia* species are confined to vacuoles within their host cells, and as suggested above, similarities in environment may have contributed to a decreased divergence in tetranucleotide frequencies.

An interesting feature of the *E. ruminantium* genome is that it contains large numbers of tandem repeats, a characteristic not found in *E. canis* or *E. chaffeensis*. These tandem repeats appear to be involved in a continuous process of genome expansion and contraction, resulting in an unusually large percentage of intergenic sequence as well as

the presence of truncated genes [107]. It has also been suggested that *E. ruminantium* is poised to undergo rapid genome rearrangements as an evolutionary strategy in the face of new environmental challenges [107; 109]. The tree of alphaproteobacteria proposed by Williams et al [122] suggests that *E. chaffeensis* and *E. canis* are sister taxa, with the strains of *E. ruminantium* being their closest relatives. The presence of truncated or chimeric genes in the *E. ruminantium* genome might help to explain the lower-than-expected average nBLASTP scores in constrast to the small G+C distances, 16S rDNA distances, and tetramer composition. It is possible that *E. ruminantium* underwent a series of rapid genome rearrangments as it initially adapted to its ruminant host, leading to a number of truncated genes without signficantly altering the G+C content and underlying tetranucleotide composition. This would be sufficient to explain the decreased BLASTP scores observed between *E. ruminantium* and the other *Ehrlichia* species, and would also explain why G+C distance, 16S rDNA distance, and tetramer Euclidean distance suggest that the genomes are in fact closely related.

Rickettsia spp.

In constrast to the *Ehrlichia* outliers examined above, the 4 *Rickettsia* positive outliers have a much higher CA than would be expected given their average nBLASTP scores (Figure 3.1). The CA values for these outliers range from 88.6% - 90.4%, whereas other genome comparisons with similar nBLASTP scores (0.80 – 0.87) range in CA from 57.46% - 77.09%. Interestingly, Figure 3.2 shows that the *Rickettsia* comparisons have the smallest 16S rDNA distances of all of the outliers considered, suggesting that the *Rickettsia* species are more closely related than the other outlier pairs, despite the higher-than-expected CA values. Figure 3.4 shows that the Rickettsia outliers have moderate differences in tetranucleotide usage; the tetramer compositions are considerably more divergent than those of the *Ehrlichia* outliers, despite the lower average nBLASTP scores associated with the *Ehrlichia* comparisons.

Although the *Rickettsia* species included in this study are obligate intracellular pathogens, they are not confined to vacuoles as are *A. phagocytophilum*, *N. sennetsu*, and the *Ehrlichia* spp. Instead, *Rickettsia* reside in the cytoplasm of an infected cell, and have

adapted to take advantage of nucleotides, amino acids, and other compounds present in the host cytosol [110]. Many *Rickettsia* lack the ability to synthesize nucleotides altogether, and have lost genes that encode or regulate a large number of biosynthetic pathways [113; 115]. The ability to depend on the host cell for amino acids might have played a role in helping to shape the tetranucleotide compositions of the various *Rickettsia* genomes, as selective pressure against metabolically expensive amino acids would be greatly reduced.

Other factors that might influence tetranucleotide variation and contribute to the positive residual CA observed for the *Rickettsia* outliers are sequence repeats and mobile elements. The *R. felis* genome contains 782 small panlindromic repeats, 85 of which were found in open reading frames, and as many as 82 genes encoding transposases [112]. Such features could easily influence the underlying tetranucleotide composition of the genome, and thus improve classification.

LGT might also have contributed to the tetranucleotide divergence observed in the *Rickettsia* outliers. Blanc et al. previously provided evidence for LGT between *R. massiliae* and *R. bellii* [117]. Although no LGT events have yet been documented for the *Rickettsia* genomes used in the present study, there is evidence for a conjugative plasmid in the genome of *R. felis* [112].

Conclusions

The results presented in this chapter demonstrate that pairwise genome distinguishability is generally proportional to 16S rDNA distance, G+C distance, and tetramer Euclidean distance, and inversely proportional to both the lowest common taxonomic rank and average nBLASTP scores. Although the CA vs. nBLASTP model is able to provide a reasonable approximation of the relationship between distinguishability and genome similarity (R^2=0.7761), it is clear from the examination of the outliers in Table 3.1 that a variety of factors may potentially influence the relative distinguishability of a given pair of genomes. Notably, similarities in both the lifestyle and environment of two organisms may affect their distinguishability. For example, all of the negative outliers

examined in this study consist of intracellular pathogens that are confined to vacuoles within mammalian host cells. Conversely, the positive outliers are either free-living aquatic bacteria, or intracellular pathogens that live freely within the cytosol of the host cell. For certain genome pairs, it appears that distinguishability may be influenced by the presence of palindromic repeats, phage DNA, and transposases in one or more of the genomes, which may result in changes to the underlying compositional patterns present in the genome. Similarly, a lack of DNA repair mechanisms or a propensity for genome rearrangement may also alter the genome signature, and therefore influence the relative distinguishability of a pair of genomes.

Although the results presented in this chapter are specific to alphaproteobacteria, the tendencies for increased or decreased distinguishability observed for certain outlier pairs are likely to be generalizable to other classes of bacteria that have similar characteristics to those discussed above. Notably, intracellular pathogens that are confined to vacuoles or similar cellular compartments may show decreased distinguishability as a result of limited opportunity for LGT and the unusual selective pressures conveyed by these environments. Conversely, cytosol-bound intracellular pathogens may show a tendency for increased distinguishability due to greatly relaxed metabolic constraints: for example, organisms that are free to make use of metabolically expensive amino acids present in the cytosol will not be under the same selective pressure as free living bacteria, and mutations resulting in an increased demand for such substrates will be more likely to persist in these populations, thus altering genome composition.

When attempting to classify two genomes on the basis of genome signature, it is apparent that the relative level of distinguishability is most limited by the compositional similarities between the two genomes. The results of this chapter highlight the fact that such compositional similarities do not necessarily correlate with similarities in phylogenetic marker genes. As such, very closely related species may prove to be highly distinguishable if various factors have caused their genome signatures to diverge. Conversely, distantly related species may show drastically decreased distinguishability if their genome signatures have converged.

Chapter 4 – Pairwise Classification of 774 Bacterial and Archaeal Genomes Based on the Tetranucleotide Profiles of Short Genomic Fragments

Motivation

The results from the preceding chapter demonstrate that the pairwise distinguishability of two α-proteobacterial genomes can be modelled using the average normalized BLASTP (nBLASTP) score of their shared orthologs. Although the relationship between average nBLASTP and classification accuracy is useful in helping to bring attention to pairs of genomes that are easier or more difficult to classify than predicted by the model, the applicability of this method is limited by its underlying dependence on the identification of orthologs. As the model is built using only orthologous sequences shared by a given pair of genomes, it essentially excludes the impact of intergenic regions and non-orthologous genes on classification.

In the present experiment, the pairwise distinguishability of genomes is measured using SVM models based upon the tetranucleotide composition of short genomic fragments rather than shared orthologs. The use of genomic fragments ensures that all regions of the given genomes are equally represented in the SVM datasets, and unlike the experiment in Chapter 3, has no dependence on gene annotations or reciprocal best BLASTP scores. Since metagenomic projects involving high-throughput sequencing ultimately generate short fragments containing mixtures of both coding and noncoding sequence, it is important to understand the degree to which such fragments may be distinguished on the basis of genome signature.

Although it has been shown that DNA fragments from a given genome tend to vary less in composition than fragments from different genomes, within-genome compositional variation is in many cases sufficient to highlight regions of putative LGT [57; 123; 124], identify genomic islands [125], or to distinguish between genes based on translational efficiency [126]. In order to identify such regions within a genome, all of

these methods rely one of several forms of clustering of the genome sequence or a representation of codon usage patterns extracted from the coding sequence. Popular clustering techniques applied to the analysis of within-genome compositional variation include adaptations of Kohonen's self-organizing map (SOM), k-means clustering, and hierarchical clustering. Given that the overall compositional signature for a given genome is actually a mosaic signature comprising multiple compositional features, it may in fact be the case that genomic regions containing these features might differ in their relative distinguishability against a comparator genome. In order to evaluate the potential differences in distinguishability exhibited by these regions, this experiment introduces k-means clustering of each genome's tetranucleotide profiles prior to SVM classification. The use of clustering will provide both the ability to compare classification accuracy on a per-cluster basis, as well as the opportunity to determine whether or not clustering of the tetranucleotide profiles enhances the SVM's ability to discriminate between genomes on the basis of their compositional signatures.

Bacterial and Archaeal genomes are typically gene dense, consisting primarily of long coding genes separated by much shorter intergenic sequences. The *E. coli* O157:H7 str. EC4115 genome, for example, contains 5,477 genes with an average length of 867 bp, representing 83.3% of the entire genome. If this genome was to be partitioned into 500 nt fragments at random, many of the fragments would contain 500 consecutive coding bases from a single gene, while other fragments would contain regions from multiple genes or a mixture of coding and non-coding sequence at varying proportions. For such hybrid fragments, the underlying compositional signatures could contribute to the degradation of performance of the SVM classifier if the result is an averaging of the tetranucleotide usage patterns for coding and noncoding sequences. The effect of hybrid fragments on SVM classification will be examined in this experiment using two measures of fragment heterogeneity: 1) the number of gene boundaries present in the fragment, and 2) the longest stretch of consecutive coding bases present in the fragment.

The goal of the present experiment is to gauge the pairwise distinguishability of 774 complete Bacterial and Archaeal genomes based on the tetranucleotide composition

of 500 nt-long fragments from each genome. As in the experiment described in Chapter 4, Support Vector Machines (SVMs) are used to build 2-class models trained on tetranucleotide frequency profiles, and the cross-validated classification accuracy is subsequently interpreted in the context of various measures of sequence similarity such as 16S rRNA distance, G+C% distance, average tetranucleotide distance, and lowest common taxonomic rank. The influences of fragment heterogeneity, the annotated biological functions of any genes encoded on a fragment, and k-means clustering of the tetranucleotide profiles on classification accuracy are also investigated.

Experimental Design

Data Acquisition and Sequence Extraction

The Genbank files for 774 complete microbial genomes were acquired from NCBI via rsync on November 28, 2008. These 774 genomes represented all of the Bacterial and Archaeal genomes available through NCBI at that time, a significant increase over the previous ortholog-based experiment in terms of both the number of genome sequences as well as the breadth of their taxonomic distribution. Whereas the previous experiment focused only on 56 genomes within the α-proteobacteria, the present experiment makes use of 721 Bacterial and 53 Archaeal genomes. In total, 472 uniquely named Bacterial species and 49 uniquely named Archaeal species are represented in the 774-genome dataset, with an average genome size of 3.58 Mbp.

In preparation for genome parameterization, the DNA sequences for all genomes were extracted directly from their respective Genbank files using a custom Perl script. In many instances, a given genome was comprised of multiple Genbank files, each representing an individual chromosome or plasmid. In such cases, each component sequence was extracted and processed individually rather than concatenating the individual sequences into a single hybrid sequence. The retention of each genome's chromosome and plasmid sequences as distinct entities throughout the experiment allows us to examine differences in distinguishability between each of the individual genome components, which may help to identify regions containing compositional biases or

sequence acquired via LGT.

Genome Parameterization

For each of the 774 genomes, all associated chromosome and plasmid sequences were partitioned into 500 nt non-overlapping fragments beginning with the first annotated position in each sequence. A fragment size of 500 nt was chosen for this experiment because previous analyses (see Chapter 2) clearly demonstrated that 500 nt fragments provide sufficient compositional signal for training SVMs to distinguish between genomes on the basis of genome signature, and furthermore, the resulting SVM models could be applied to fragments that are greater than 500 nt in length. Although other projects such as PhyloPythia [38], TACOA [37], and tetra-ESOM [127] chose to use longer fragment sizes (800 nt, 1000 nt, 5000 nt respectively) in their analyses, there are several notable advantages to using a shorter fragment size: 1) both metagenomics and next-generation sequencing are generating datasets that contain short DNA sequences well below 800 nt in length, and accurate methods for binning such sequences do not presently exist; 2) many microbial genomes, especially those of obligate endosymbionts, are relatively small and will generate only a limited amount of SVM training data with larger fragment sizes; and 3) if the genome distinguishability results are to be applied to the identification of putative instances of LGT, shorter fragment sizes will provide greater resolution for identifying the specific regions suspected to be involved in a given LGT event.

Within-genome variation of genome signature tends to decrease as fragment size increases (Chapter 2), and as a result, larger fragment sizes will usually produce higher classification accuracies than shorter fragment sizes using the same data set. Although this would appear to support the use of longer fragments whenever possible, the increase in classification accuracy is likely a side effect of oversimplifying the underlying classification problem. Additionally, it was demonstrated in Chapter 2 that SVM models trained using 500 nt genomic fragments are able to classify longer fragments with a high degree of accuracy, whereas models trained using longer fragments exhibited reduced classification accuracies when confronted with shorter fragments.

For each DNA fragment defined in the previous step, the frequency of each of the 256 possible tetranucleotides was calculated using a sliding-window approach (step = 1). For a fragment of length B (where B >= 4) there are B-3 possible overlapping tetranucleotide positions. The tetranucleotide profile for each fragment was calculated by iterating over each of the B-3 windows along the coding strand, and incrementing a counter for each of the 256 possible tetranucleotides as they were encountered in the DNA sequence. The result is a 256-element vector containing the frequencies of each of the tetranucleotides in a fragment. Since not all fragments were 500 nt in length (i.e., the last fragment in a chromosome/plasmid is often less than 500 nt) it was necessary to normalize the frequencies by dividing each frequency vector by the length of the fragment.

Previous studies have adjusted for strand and G+C biases in oligonucleotide frequency data by 'symmetrizing' the oligonucleotide frequencies and correcting the frequencies based on local G+C content, respectively [80; 128]. In addition to the unsymmetrized frequencies calculated above, a set of symmetrized tetranucleotide frequency vectors were determined for all genomes. Symmetrized tetranucleotide frequencies were calculated in the following manner: tetranucleotide counts were first calculated for both the coding and template strands using the sliding-window approach described above. The resulting set of tetranucleotide counts was then reduced to the set of non-redundant tetranucleotide counts by combining the count for each tetranucleotide with the corresponding count of its reverse-complementary tetranucleotide and dividing by 2. The set of 136 non-redundant tetranucleotide counts were then converted to non-redundant (symmetrized) tetranucleotide frequencies by dividing each count by the length of the associated fragment.

In order to correct for G+C content, the symmetrized tetranucleotide frequencies for each fragment were adjusted using the following formula:

$$G = \log_2\left(\frac{S_t}{\frac{1}{2}(fn_1\, fn_2\, fn_3\, fn_4)}\right)$$

where S_t represents the symmetrized tetranucleotide frequency for a particular tetranucleotide *t* in the given fragment, and fn_1 through fn_4 represent the symmetrized mononucleotide frequencies of each of the component nucleotides in *t* as determined by the symmetrized G+C for the fragment. An additional set of symmetrized and G+C-corrected tetranucleotide frequencies were calculated in an identical manner, with the exception that the frequencies of each of the component mononucleotides fn_1 through fn_4 were based upon the symmetrized G+C content for the entire source genome as opposed to the local symmetrized G+C for the 500 nt fragment. The initial set of 299,151 pairwise SVM trials in this study were performed separately using both the unsymmetrized and symmetrized tetranucleotide frequency profiles in order to examine the influence of symmetrization on classification accuracy. Additionally, 500 SVM-based comparisons between randomly selected pairs of genomes were performed for both sets of G+C-corrected tetranucleotide frequency profiles (fragment-based or genome-based) to likewise gauge the impact of G+C correction on distinguishability.

Measuring Pairwise Distinguishability Using Support Vector Machines

SVMs were used to quantify the distinguishability for each of the 299,151 possible pairwise comparisons among the 774-genome dataset. For a given genome pair, all tetranucleotide frequency profiles associated with each of the two genomes were first compiled into a single SVM data file. This large SVM data file was subsequently split into 5 cross-validation (CV) groups using a random stratified assignment algorithm to maintain consistent class representation among the 5 CV groups. Grid searches were performed on random 500-item subsets of the CV groups in order to determine appropriate values for C and γ (explained in Ch.3). Lastly, SVM models were built and tested via libSVM v2.88 using a 5-fold leave-one-out cross-validation scheme, and the classification accuracy was recorded. As in the previous experiments, the Gaussian/RBF kernel function was selected as it was shown to outperform the linear kernel for oligonucleotide frequency datasets [38].

In addition to the overall classification accuracy for a given pair of genomes, the

individual classification (correct/incorrect) of each individual fragment involved in the comparison was also recorded for use in subsequent analyses.

Two post-processing steps were required in order to resolve inconsistencies in the SVM classification results. Of the 299,151 possible genome pairs, 406 (0.14%) reported classification accuracies less than 50%; a paradoxical result for a 2-class classification problem. Further investigation revealed that all affected pairs involved conspecific organisms, which were expected to give approximately 50% classification accuracy since the genomes involved in these pairs were nearly identical in all cases. Repeat runs of the affected pairs did not resolve the sub-50% CAs. It is possible that libSVM was unable to correctly handle these instances of essentially unclassifiable training sets, leading to unrealistic CAs in this small number of cases. As a solution to this issue, the CA for the 406 affected pairs was set to exactly 50% prior to including the results in subsequent analyses.

The second inconsistency in the SVM results affected 17 (0.0057%) of the 299,151 pairs. For these 17 pairs, the reported overall classification accuracy was 50% despite the fact that the grid search CA for these same pairs was always 97% or greater. It is likely that the heuristic grid search failed to choose reasonable values for C and γ in these cases, and as a result, the SVM incorrectly classified the total complement of fragments in the affected pairs as one genome or the other, leading to a CA of 50%. In order to correct for these inconsistencies, all SVM runs with an overall CA at least 3% less than the reported grid search CA (31 pairs in total) were repeated. Of the 31 re-runs, the 17 inconsistent pairs no longer reported inconsistent CAs, and the remaining 18 pairs showed little or no change in overall CA.

Outlier Comparison

Once the pairwise classification accuracies were determined for all 774 genomes, the classification results were interpreted in relation to a number of measures of sequence similarity: difference in genomic G+C, 16S rDNA distance, lowest common taxonomic rank, and the difference in average genomic tetranucleotide composition. Wherever

possible, models were constructed from the resulting plots and their statistical significance was evaluated. Additionally, a number of positive and negative outliers were identified and selected for inclusion in subsequent analysis pipelines.

Difference in Genomic G+C Content

For each genome, the total G+C content was calculated as the total number of G and C nucleotides in all chromosomes and plasmids divided by the total number of nucleotides in all of these sequences. Once the genomic G+C values were calculated, the classification accuracy of each pair of genomes was plotted against the difference in G+C for the given genomes.

16S rDNA Distance

The Ribosomal Database Project (RDP) Release 10.10 was queried using the RefSeq accession numbers associated with all 774 genomes in order to compile a list of relevant 16S rDNA sequences. Although several of the genomes could not be mapped to RDP sequences using this method, a total of 706 bacterial and 43 archaeal 16S rDNA sequences were queried successfully. The myRDP interface of the RDP project was subsequently used to generate uncorrected distance matrices for the given 16S rDNA sequences, and for each pair of genomes, the 16S rDNA distance was extracted and/or calculated from these matrices.

A total of 646 genomes were associated with only a single 16S rDNA sequence in the RDP, and as such, the pairwise 16S rDNA distance for any two such genomes could easily be extracted directly from the myRDP-generated distance matrices. A small number of genomes, however, were associated with multiple 16S sequences in the RDP: 35 genomes were linked to 2 16S sequences, while 12 genomes contained exactly 3 entries in the RDP. For pairwise comparisons in which one or both of the genomes contained multiple 16S rDNA entries in the RDP, the average between-genome 16S rDNA distance was calculated using all 16S rDNA sequences associated with each genome. Classification accuracy was plotted against the set of 16S rDNA distances, and an exponential model was fit using the R Statistical Computing Package.

One caveat to the use of RDP is that the myRDP interface is unable to provide a distance matrix comparing bacteria vs. archaeal 16S sequences and as a result, no Bacteria vs. Archaea comparisons are present in the classification accuracy vs. 16S rRNA distance plot or model.

Lowest Common Taxonomic Rank

For each pair of genomes, the most specific taxonomic rank shared by both genomes (lowest common taxonomic rank) was determined and the set of such values was utilized in order to partition the CA vs. genomic G+C plot in terms of taxonomy. This partitioning allows the boundaries of distinguishability to be qualitatively examined in terms of the taxonomic relatedness of the organisms in question.

Difference in Average Tetranucleotide Composition

For each genome, the average tetranucleotide profile was calculated by summing the individual tetranucleotide counts across all fragments in the genome, and then dividing the set of tetramer counts by the total number of fragments. Pairwise tetranucleotide distance was then calculated as the Euclidean distance between the average tetranucleotide compositions of each pair of genomes. Classification accuracy was plotted against average tetranucleotide composition, and R was used in order to fit an exponential model.

Evaluating the Impact of Composition-based Clustering, Fragment Heterogeneity, and Fragment Functional Annotations on Classification

A subset of 16 genome pairs from the CA vs. 16S rDNA distance plot were selected for analysis using 3 additional pipelines (Table 4.1). A variety of genome pairs were selected on the basis of their residual values from the fitted model or other interesting properties of the pairs, for example congeners that have higher than expected CA, or distantly related organisms that have less than expected CA. Other genome pairs were selected in order to include pairs that have CA values in each of the ranges 55% - 60%, 60% - 70%, 70% - 80%, 80% - 90%, and 90% - 100%.

Table 4.1: Outliers Selected for Inclusion in K-means Clustering, Fragment Heterogeneity, and Functional Profiling Pipelines

Each of the following genome pairs was selected for inclusion in the outlier analysis pipeline. CA indicates the classification accuracy for each genome pair during the initial 299,151 SVM trials. LCTR denotes the lowest common taxonomic rank shared by each pair. Residuals are based on the CA vs. 16S rDNA distance model.

Genome1	Genome2	CA	LCTR	Residual
Methanosarcina barkeri str. Fusaro	*Gramella forsetii*	89.59%	None	-0.10
Ehrlichia ruminantium	*Methanosphaera stadtmanae*	88.89%	None	-0.11
Pyrococcus abyssi GE5	*Metallosphaera sedula* DSM 5348	91.93%	Domain	-0.07
Buchnera aphidicola str. Cc (Cinara cedri)	*Candidatus Sulcia muelleri* GWSS	83.92%	Domain	-0.15
Prochlorococcus marinus	*Borrelia afzelii*	84.65%	Domain	-0.14
Chlamydophila abortus S26/3	*Neorickettsia sennetsu* str. Miyayama	85.70%	Domain	-0.13
Prochlorococcus marinus str. AS9601	*Candidatus Pelagibacter ubique* HTCC1062	86.44%	Domain	-0.12
Bradyrhizobium japonicum USDA 110	*Mesorhizobium loti* MAFF303099	80.95%	Order	-0.15
Lactobacillus acidophilus NCFM	*Lactobacillus gasseri* ATCC 33323	66.77%	Genus	-0.25
Haloarcula marismortui ATCC 43049	*Halobacterium salinarum* R1	87.31%	Family	-0.09
Prochlorococcus marinus str. MIT 9303	*Prochlorococcus marinus* str. AS9601	97.40%	Species	0.12
Haemophilus somnus 2336	*Pediococcus pentosaceus* ATCC 25745	93.43%	Domain	-0.05
Borrelia duttonii Ly	*Borrelia recurrentis* A1	55.94%	Genus	0.06
Nitrobacter hamburgensis X14	*Nitrobacter winogradskyi* Nb-255	68.44%	Genus	-0.01
Shewanella baltica OS195	*Shewanella denitrificans* OS217	78.67%	Genus	-0.06
Nitrosospira multiformis ATCC 25196	*Nitrosomonas eutropha* C91	89.52%	Family	-0.03

K-means Clustering

The normalized tetranucleotide frequency vectors for each of the selected genome pairs were independently clustered using the *kmeans* method provided by R v2.8.1 for $k \in \{2, 3, 4, 5, 6\}$, where k represents the number of clusters. At each value of *k*, the *2k* cluster assignments for each outlier pair were used to designate class labels in the corresponding SVM training file. In total, 6 SVM training files were generated for each outlier pair; one for each of the 5 values of *k* utilized in the k-means clustering step, plus a control case where no clustering was used (essentially, $k = 1$).

Grid searches were performed on 1000-element subsets of each of the SVM training files in order to determine reasonable values of C and γ, and SVM models were subsequently trained and evaluated using 5-fold, leave-one-out cross validation as previously described. 1000-element subsets were used in the present grid searches (as opposed to 500-element subsets used in the larger set of SVM trials) in order to help reduce the likelihood that inappropriate C and γ values might be selected. Two classification accuracies were recorded for each SVM model: a strict classification accuracy in which correct classification was defined as the SVM's ability to correctly predict a given fragment's cluster assignment, and a relaxed classification accuracy in which correct classification was defined as the SVM's ability to correctly predict a given fragment's source genome, regardless of whether the the fragment was assigned to the correct cluster.

In an attempt to understand the specific compositional features that determine the assignment of a genomic fragment to a given cluster, two additional analyses were performed following the k-means clustering step. In the first analysis, the total number of plus strand and minus strand coding nucleotides that fall within each fragment were determined using the gene coordinates in the respective Genbank files. Next, the proportions of coding nucleotides were aggregated by cluster ID in order to determine the overall distribution of plus strand and minus strand coding nucleotides for each cluster. In the second analysis, the average G+C content for each cluster was calculated by determining the total number of G and C nucleotides in all fragments assigned to each

cluster, and then dividing by the total number of nucleotides in all fragments assigned to the same cluster.

Fragment Heterogeneity

Each of the individual nucleotides within a Bacterial and Archaeal genome can belong to one of two general classes of sequence: 1) protein-coding sequences (CDS), which contain all nucleotides that fall within one or more open reading frames, and 2) intergenic sequences (IGS), which represent all of the non-protein-coding nucleotides that exists between open reading frames. In the present study, fragments from each genome were analyzed using two measures of fragment heterogeneity: 1) the total number of sequence boundaries present in the fragment, and 2) the longest contiguous block of coding nucleotides present in the fragment. Two basic types of sequence boundaries may exist in a given fragment: 1) *CDS* → *CDS* transitions occur between adjacent open reading frames that lack intervening intergenic sequence, and 2) *CDS* → *IGS* (and similarly, *IGS* → *CDS*) transitions occur between adjacent open reading frames and neighboring intergenic sequence. Mann-Whitney tests were performed using R in order to determine whether correctly classified fragments were more or less heterogeneous than incorrectly classified fragments for each of the outlier pairs in Table 4.1.

Fragment Functional Annotations

The functional annotations for each fragment were examined in order to test whether or not fragments from certain functional classes are easier or more difficult to classify. For each fragment, the distributions of specific TIGR main roles and sub roles for annotated genes that overlap the given fragment were determined. In some cases, 2 or more genes may overlap a given fragment, in which case the fragment may have several associated TIGR main roles or sub roles.

Examination of the TIGR main role aggregate data indicated unnecessary redundancy in a number of the main role categories. As such, post-processing of the TIGR data was performed in order to consolidate several of the 'unknown' and 'hypothetical protein' categories, for example, by merging the 'hypothetical protein' and

'hypothetical proteins' categories into a single TIGR main role.

Chi-squared tests were performed using R in order to examine whether or not the difference in distribution of correctly classified and incorrectly classified fragments into the various TIGR functional categories are statistically significant.

Investigating Convergence of Genome Composition and Putative LGT

Three genome pairs (*Methanosphaera stadtmanae* vs. *Ehrichia ruminantium* str. Welgevonden, *Prochlorococcus marinus AS9601* vs. *Pelagibacter ubique*, and *Haloarcula marismortui* vs. *Halobacterium salinarum*) were chosen in order to search for possible instances of LGT or convergence in genome composition. *M. stadtmanae* vs. *E. ruminantium* str. Welgevonden represents a comparison between parasitic Archaeal and Bacterial species that have lower than expected CA as predicted by the CA vs. 16S rDNA best fit model. One possible hypothesis is that these distantly related organisms have undergone recent LGT, resulting in portions of one or both genomes that have unameliorated genome signatures [55]. In such an instance, a significant portion of unameliorated sequence could reduce the resulting CA of the SVM classifier. Similarly, the decreased CA observed for the halophiles *Haloarcula marismortui* vs. *Halobacterium salinarum* may also be explained by the same hypothesis of a recent LGT event (or series of LGT events).

P. marinus AS9601 vs. *P. ubique* represents a comparison between two marine Bacterial species from the phyla Cyanobacteria and Proteobacteria, respectively. This genome pair exhibited a lower than expected CA of 86.44% according to the CA vs. 16S rDNA model. One hypothesis for this genome pair is that convergence in genome composition due to the reduced nature of the genomes and similarities in niche have resulted in decreased CA.

Correct and Incorrect Fragment Classification Versus Genome Position

In cases of recent unameliorated LGT, such sequences may exist as regions

containing a high density of misclassified fragments along one or both of the genomes involved in the comparison. In order to identify such regions of misclassification, all chromosomes and plasmids from each genome were recoded as binary sequences representing correct/incorrect classifications for all 500 nt fragments contained within the genome. The resulting binary sequences were analyzed to find intervals containing at least 37.5% misclassified fragments. This minimum of 37.5% misclassified fragments was chosen because smaller cut-offs tended to result in the identification of regions of misclassification that included long stretches of correctly classified fragments. Circos [129] was subsequently used to plot both the binary classification sequences and the identified intervals of misclassification for the 3 genome pairs.

Distribution of nBLASTP Values for Orthologs Contained Within Misclassified Fragments

In the event that recent LGT might be contributing to reduced CA for a given genome pair, one might reasonably expect that lack of amelioration could result in LGT-derived orthologous genes having higher normalized BLASTP (nBLASTP) scores as compared to orthologous genes acquired through ancient LGT or sequences inherited vertically from the most recent common ancestor. In order to compare the nBLASTP scores of correctly versus incorrectly classified fragments, the reciprocal best hit method was used to query all orthologous pairs of genes for each genome pair. The total set of orthologs for each pair of genomes was partitioned into 'correct' and 'incorrect' bins depending on the location of each ortholog relative to the previously identified regions of misclassification in one or both of the genomes. If 95% or more of the nucleotides in a given ortholog overlap with a region of misclassification in either genome, the given ortholog is considered to be incorrectly classified, while all remaining orthologs are considered to be correctly classified. For each genome pair, the average nBLASTP scores were determined for all correctly classified orthologs, incorrectly classified orthologs, as well as the complete set of orthologs. Histograms and population density distribution plots for the resulting nBLASTP scores were subsequently generated and the two-sided Mann-Whitney test was used to compare the nBLASTP distributions using R.

Results

2-class SVMs were used to train models for all possible pairings of 774 bacterial and archael genomes, and pairwise distinguishability was calculated as the classification accuracy of each SVM model using a 5-fold cross validation approach. In general, the majority of genome pairs were highly distinguishable, with 93% of the 299,151 pairings leading to a classification accuracy of 95% or greater (mean: 98.1%). Few genome pairs showed classification at or slightly above baseline, with 0.0029% of comparisons leading to classification accuracies of 55% or less.

Influence of Tetranucleotide Symmetrization and G+C Correction on Classification

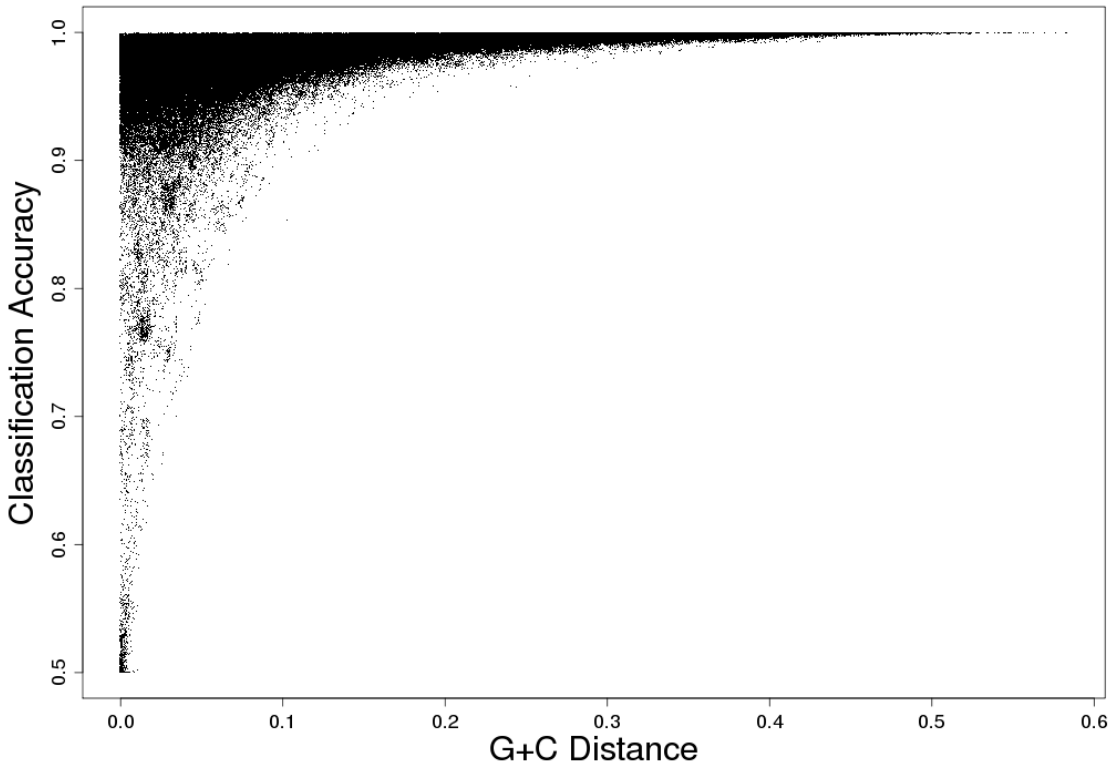
In the vast majority of cases, symmetrization of tetranucleotide frequencies and correction for G+C content had little impact on classification accuracy. Of the 298,589 pairs of genomes that demonstrated CA >= 50%, over 99.5% showed a difference in CA of <1% between the unsymmetrized and symmetrized SVM runs, while the difference in means was only 0.06% (98.20% vs. 98.26%). Of the 0.5% of trials that gave a difference in CA of >1%, symmetrization resulted in an increase in CA for 901 genome pairs, while a decrease in CA was observed for the remaining 526 cases.

Of the 500 genome pairs examined using tetranucleotide frequency profiles that were both symmetrized and corrected for local fragment-based G+C content, the 435 cases with CA >=50% demonstrated a close fit to the non- G+C corrected runs ($y = 0.98x + 0.497$; $R^2 = 0.99$). Additionally, a paired-sample t-test indicated a significant decrease in CA when using the local G+C correction, with $p = 3.2 \times 10^{-26}$.

A statistically significant increase in CA was observed for the subset of SVM runs performed using symmetrized tetranucleotide frequencies that were corrected for genome-level G+C content. A paired-sample t-test indicated a *p*-value of 1.7×10^{-8}. Although this is an interesting result, correcting for genome-level G+C content is of little practical value to the classification of short anonymous DNA fragments, as it necessitates the availability of complete genome sequences for both genomes prior to classification as

Figure 4.1: Classification Accuracy Versus Genomic G+C Distance

Classification accuracy was plotted in terms of genomic G+C distance for all pairs of genomes. The G+C content for each genome was calculated as the total number of G/C nucleotides in the genome (including all chromosomes and plasmids) divided by the total number of nucleotides.



well as *a priori* association of each fragment to its source genome.

G+C Distance

G+C distance imposes a lower bound on CA, with instances of CA >= 99.9% being observed for comparisons across the full spectrum of G+C distances (Figure 4.1). For the 62,610 genome pairs with up to a 5% G+C distance, CA ranges from 50% - 99.97%, with a corresponding mean CA of 95.1%. Within this set of comparisons, the low G+C distances might contribute to the convergence of genome signatures for some genome pairs, although it is important to note that even small G+C distances allow sufficient variability in tetranucleotide composition for certain genome pairs to be distinguished with nearly 100% CA. The 54,375 genome pairs with G+C distances in the range 5-10% have CA values between 78.75% - 99.98% (mean CA = 97.59%). Pairs with a G+C distance of at least 10% are highly distinguishable, as indicated by a minimum observed CA of 85.32% for the 182,166 comparisons that fall into this category.

Tetranucleotide Euclidean Distance

Both minimum and maximum bounds on CA are observed when CA is plotted in terms of tetramer Euclidean distance (Figure 4.2). For tetramer distances less than 0.5%, CA ranges from 50% - 72.95% with a mean CA of 53.43%. As tetramer distance increases from 0.5% - 2.5%, CA varies approximately linearly with a mean CA of 94.6% (minimum CA = 52.47%, maximum CA = 99.63%). The vast majority of genome pairs (251,676) have tetramer distances greater than 2.5% and are almost completely distinguishable. Within this set of comparisons, CA ranges from 84.64% - 100%, with a mean CA of 98.96%. The 14,820 genome pairs with at least a 10% tetramer distance have a mean CA of 99.87%, with pair no exhibiting a CA of less than 98.85%. The best-fit exponential model results in an R^2 of 0.8422 (*p*-value: $<2.2e^{-16}$).

16S rDNA Distance

CA is proportional to 16S rDNA distance as illustrated in Figure 4.3. For genome pairs with little to no 16S distance (0% - 0.5%), CA ranges from 50% - 79.42% (mean

Figure 4.2: Classification Accuracy Versus Average Tetramer Euclidean Distance

Classification accuracy was plotted with respect to the tetramer Euclidean distance for each genome pair (grey dots). The solid line represents the best-fit exponential model (R^2 = 0.8422, p-value < 2.2e-16).

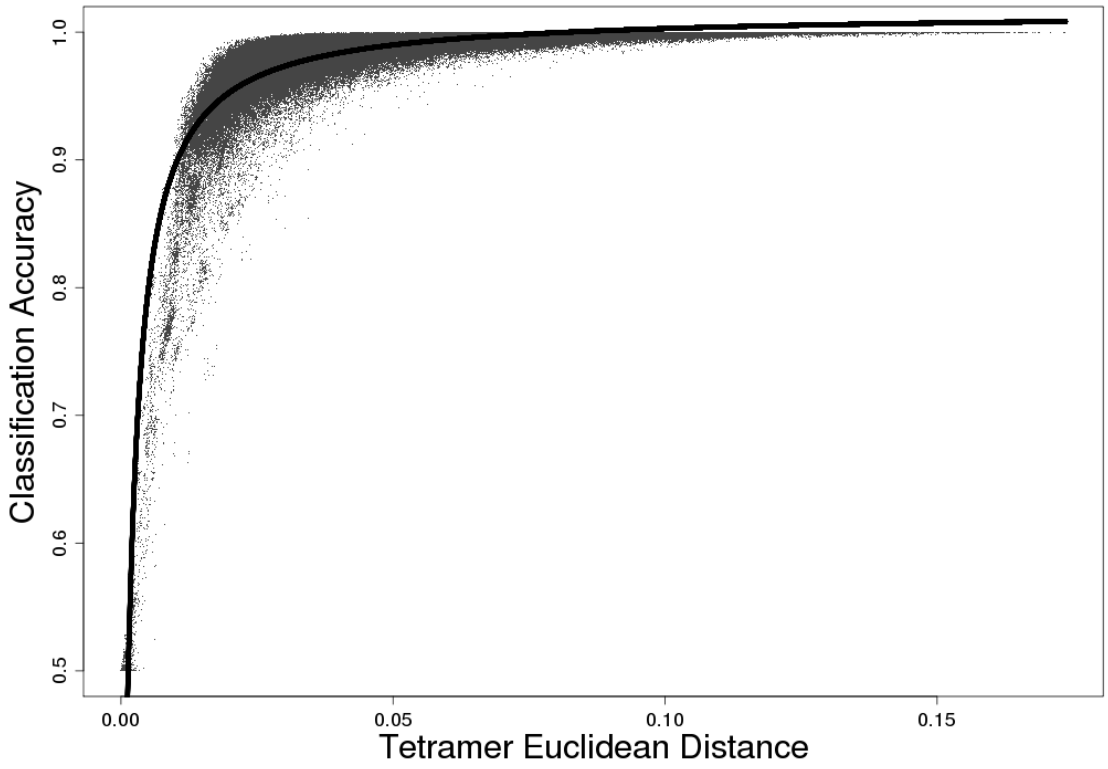
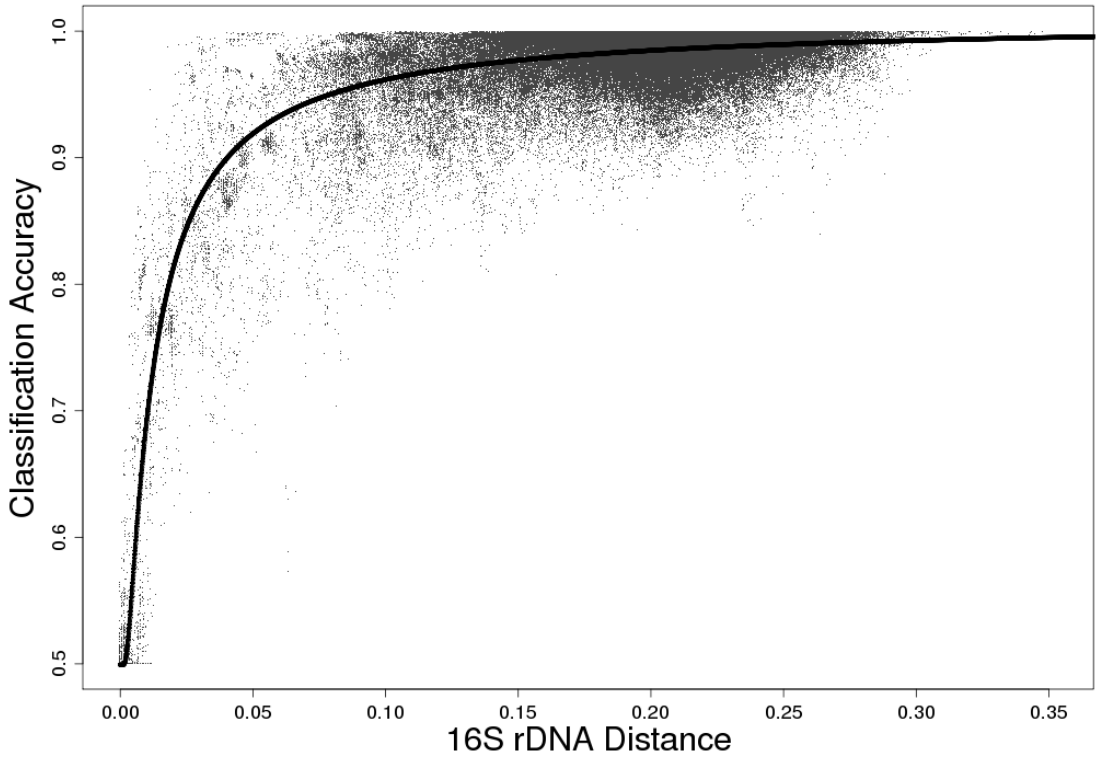


Figure 4.3: Classification Accuracy Versus 16S rDNA Distance

Classification accuracy plotted in terms of 16S rDNA distance for each genome pair (grey dots). The solid line represents the best-fit exponential model (R^2 = 0.7406, p-value < 2.2e-16).



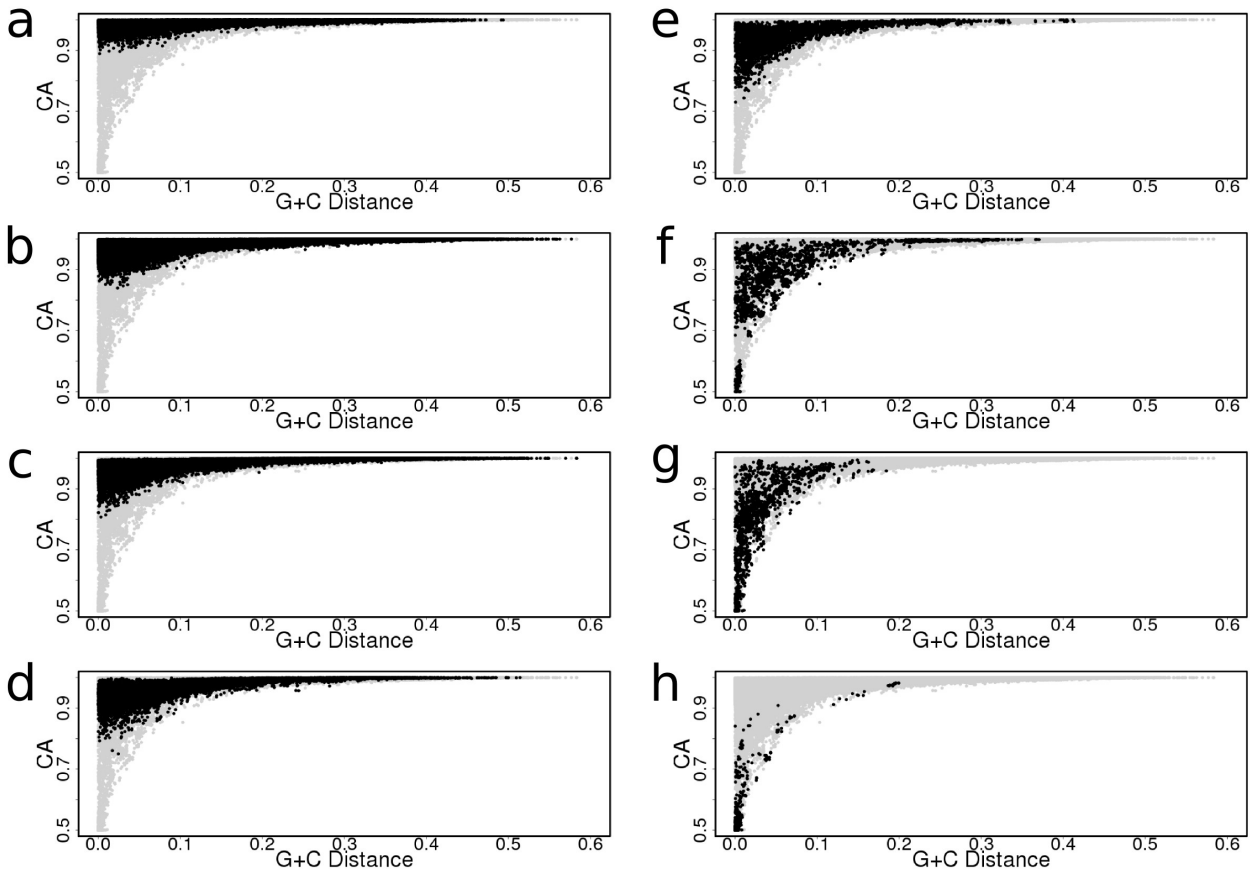
CA = 52.33%). Small increases in 16S distance quickly lead to a large range of CA values. For example, CA of >95% is achievable with 16S distances as low as 1.14% (*Mycoplasma genitalium* G37 vs. *Mycoplasma pneumoniae* M129), although the mean CA for the 83 genome pairs with similar 16S distances (1.13% - 1.15%) is only 76.48%. Furthermore, the minimum CA observed within this set of 83 genome pairs is 55.48% (*Thermoanaerobacter pseudethanolicus* ATCC 33223 vs. *Thermoanaerobacter* sp. X514). Genome pairs with 16S distances from 5% - 15% exhibit significant variability in CA, with CA values ranging from 57.34% - 99.99% (mean CA = 96.17). Above a 16S distance of 30%, genome pairs show much less variability in CA, and CA converges toward 100% (mean CA = 99.72%, minimum CA = 98.02%, maximum CA = 100%). An R^2 of 0.7406 (p-value: $< 2.2e^{-16}$) is achieved using the best-fit exponential model.

CA in Terms of the Taxonomic Relatedness of Genome Pairs

The taxonomic relatedness of genome pairs is sufficient to predict upper and lower bounds on CA (Figure 4.4). For comparisons between an archaeal species versus a bacterial species, the minimum observed CA is 88.89% (Figure 4.4a) with the majority of such comparisons resulting in near-perfect classification (mean CA = 99.07%, maximum CA = 100%). As the taxonomic ranks of two genomes become more similar, CA tends to decrease. For example, genome pairs from the same class but different orders (Figure 4.4d) have CA in the range 74.99% - 100% (mean CA = 97%), while pairs from the same family but different genus (Figure 4.4f) range in CA from 50% - 99.89% (mean CA = 88.18%). Genome pairs comprising different strains of the same species (Figure 4.4h) show the lowest overall CA (mean CA = 55.04%, minimum CA = 50%) although even at this level, 13 pairs of 9 unique strains of *Prochlorococcus marinus* show CA in excess of 95%. The next highest within-species comparisons excluding those of *P. marinus* are strains of *Buchnera aphidicola* (CA = 90.82%), *Pseudomonas fluorescens* (CA = 88.01%), and *Pseudomonas aeruginosa* (CA = 86.36%).

Figure 4.4: Classification Accuracy Versus Genomic G+C Distance Partitioned by Lowest Common Taxonomic Rank

Classification accuracy plotted in terms of genomic G+C distance for all genome pairs. The individual panels represent the pairwise comparisons with lowest common taxonomic rank at each level. Panel a) genome pairs consisting of an archaeal species vs. a bacterial species, b) kingdom, c) phylum, d) class, e) order, f) family, g) genus, h) species.



CA in Terms of the Functional Annotations Associated with Each Fragment

Genomic fragments that overlap genes assigned to specific TIGR main role categories show statistically significant tendencies for correct or incorrect classifications (Table 4.2). Across all comparisons within the 16 genome pairs listed in Table 4.1, the overall distribution of correct and incorrect fragment classifications into the various TIGR main role categories results in an overall X^2 of 42.87 (df = 17, p-value = 0.0005). Role categories that are associated with an over-representation of correct classifications include roles involved in biosynthetic pathways, such as “biosynthesis of cofactors, prosthetic groups, and carriers” and “amino acid biosynthesis”. Role categories that demonstrate a tendency toward misclassification include several roles associated with a subset of informational genes (“cellular processes”, “protein synthesis”, “signal transduction”) as well as roles associated with mobile elements (“mobile and extrachromosomal element functions”). When the analysis was repeated using the more specific TIGR subrole categories, the distribution of correct/incorrect classifications into the various sub role categories was not statistically significant (X^2 = 97.0, df = 94, p-value = 0.394).

Fragment Heterogeneity

Of the 16 genome pairs examined in the outlier analyses, 9 pairs demonstrated a significant difference in both measures of heterogeneity for correctly classified versus incorrectly classified fragments (dark grey shading, Table 4.3). For these 9 pairs, incorrectly classified fragments tended to have more gene-gene or gene-intergenic region boundaries as well as shorter coding sequences relative to their correctly classified counterparts. An additional genome pair, *Buchnera aphidicola* str. Cinara cedri vs. Candidatus *Sulcia muelleri* GWSS, exhibited an identical trend for both the number of boundaries and longest contiguous coding region, although only the difference in distribution of the boundaries showed statistical significance. Three genome pairs showed the opposite effect: *Ehrlichia ruminantium* str. Welgevonden v2 vs. *Methanosphaera stadtmanae* DSM 3091, *Prochlorococcus marinus* str. MIT 9303 vs. *Prochlorococcus*

Table 4.2: Functional Profiling and TIGR Main Role X^2 Results

Chi-square results are shown for each of the TIGR main role categories, as determined from the number of correctly/incorrectly classified fragments observed for each category. 'Correct' and 'Incorrect' denote the number of correctly or incorrectly classified fragments for each category. 'Expected' refers to the number of fragments that are expected to be correctly classified based upon the overall CA for the complete set of fragments. 'X^2' shows the corresponding Chi-square score for each TIGR main role category. 'Trend' indicates whether the Chi-square test suggests an over-representation (+) or under-representation (-) of the correctly classified fragments for the associated role category.

TIGR main role	Correct	Incorrect	Expected	X^2	Trend
Biosynthesis of cofactors, prosthetic groups, and carriers	4220	574	4015.81	10.38	+
Cellular processes	1977	538	2106.75	7.99	-
Mobile and extrachromosomal element functions	887	260	960.81	5.67	-
Amino acid biosynthesis	3517	527	3387.56	4.95	+
Protein synthesis	7292	1641	7482.95	4.87	-
Signal transduction	1194	297	1248.97	2.42	-
Hypothetical proteins	1614	245	1557.24	2.07	+
Transcription	1441	342	1493.57	1.85	-
Regulatory functions	1865	322	1831.99	0.59	+
Protein fate	3670	765	3715.09	0.55	-
Energy metabolism	6048	1104	5991.05	0.54	+
Cell envelope	2473	517	2504.65	0.4	-
Purines, pyrimidines, nucleosides, and nucleotides	2278	411	2252.51	0.29	+
Transport and binding proteins	4943	924	4914.64	0.16	+
Unknown function	12437	2375	12407.64	0.07	+
Central intermediary metabolism	1038	193	1031.18	0.05	+
Fatty acid and phospholipid metabolism	835	158	831.81	0.01	+
DNA metabolism	4883	940	4877.78	0.01	+

Table 4.3: Results of 2-sided Mann-Whitney Test of the Distributions of Fragment Heterogeneity for Correctly Classified Versus Incorrectly Classified Fragments.

'B' is used to signify values associated with the average number of gene-gene or gene-intergenic boundaries in correctly/incorrectly classified fragments. 'N' denotes values associated with the average number of nucleotides in the longest contiguous coding region in correctly/incorrectly classified fragments. Dark grey shading is used to highlight comparisons that show significantly higher fragment homogeneity in correctly classified fragments as compared to incorrectly classified fragments. Light grey shading indicates comparisons that show significantly higher fragment heterogeneity in correctly classified fragments.

Genome Pair	$B_{correct}$	$B_{incorrect}$	$B_{p\text{-}value}$	$N_{correct}$	$N_{incorrect}$	$N_{p\text{-}value}$
Methanosarcina barkeri str. Fusaro vs. *Gramella forsetii* KT0803	0.757	0.947	< 2.2e-16	427.02	402.98	< 2.2e-16
Ehrlichia ruminantium str. Welgevonden v2 vs. *Methanosphaera stadtmanae* DSM 3091	0.731	0.645	0.0207	429.02	437.19	0.02534
Buchnera aphidicola str. Cc (Cinara cedri) vs. *Candidatus Sulcia muelleri* GWSS	0.792	0.975	0.03071	428.81	418.34	0.1135
Prochlorococcus marinus str. MIT 9515 vs. *Borrelia afzelii* PKo	0.959	1.002	0.3602	419.65	413.54	0.1400
Chlamydophila abortus S26/3 vs. *Neorickettsia sennetsu* str. Miyayama	0.821	0.937	0.0071	427.26	414.29	0.003632
Prochlorococcus marinus str. AS9601 vs. *Candidatus Pelagibacter ubique* HTCC1062	0.929	1.101	5.955e-05	422.90	409.95	0.0004229
Bradyrhizobium japonicum USDA 110 vs. *Mesorhizobium loti* MAFF303099	0.840	0.870	0.008987	422.16	415.82	2.871e-06
Lactobacillus acidophilus NCFM vs. *Lactobacillus gasseri* ATCC 33323	0.828	0.895	0.003472	428.20	421.91	0.001834
Haloarcula marismortui ATCC 43049 vs. *Halobacterium salinarum* R1	0.907	0.963	0.0363	419.41	407.94	8.276e-06
Prochlorococcus marinus str. MIT 9303 vs. *Prochlorococcus marinus* str. AS9601	1.015	0.662	3.850e-06	412.94	431.11	0.002699
Haemophilus somnus 2336 vs. *Pediococcus pentosaceus* ATCC 25745	0.842	0.930	0.03006	426.50	414.62	0.002637
Borrelia duttonii Ly vs. *Borrelia recurrentis* A1	0.731	0.699	0.06359	431.52	436.70	0.01709
Nitrobacter hamburgensis X14 vs. *Nitrobacter winogradskyi* Nb-255	0.814	0.817	0.9958	418.62	418.94	0.8241
Shewanella baltica OS195 vs. *Shewanella denitrificans* OS217	0.772	0.818	0.001781	427.32	420.19	4.648e-06
Nitrosospira multiformis ATCC 25196 vs. *Nitrosomonas eutropha* C91	0.823	0.829	0.8466	424.09	419.47	0.09033
Pyrococcus abyssi GE5 vs. *Metallosphaera sedula* DSM 5348	0.835	0.959	0.004138	424.45	412.98	0.00539

marinus str. AS9601, and *Borrelia duttonii* Ly vs. *Borrelia recurrentis* A1 all showed significantly fewer boundaries and shorter coding sequences in *incorrectly* classified fragments (except for the boundary measure for the *Borrelia* comparison, with p = 0.06359). For the *Borrelia* pair, both genomes have numerous plasmids (*B. duttonii*: 16, *B. recurrentis*: 7) that contain many short genes relative to those found on the primary chromosomes. Similarly, in the closely related *Prochlorococcus* pair, approximately 50% of the genes in each genome are shorter than 250bp in length. Given this propensity for short genes, there is an increased likelihood that any randomly selected 500 nt fragment will contain a mixture of both coding and intergenic sequences. For these hybrid fragments, the faster-evolving intergenic sequences may provide a stronger genome signature on which the SVM can base its classifications, leading to the increase in classification accuracy observed for the heterogeneous fragments in both the *Borrelia* and *Prochlorococcus* pairs.

Impact of Unsupervised K-means Clustering on CA

Clustering of the tetranucleotide frequencies prior to classification degrades SVM performance (Table 4.4). In comparison to the baseline CA for each pair of genomes, clustering the tetranucleotide profiles using increasing values of *k* only serves to decrease classification accuracy. Even when the conditions are relaxed and correct classification is defined as the assignment of a fragment to the correct source genome without regard to cluster assignment within the genome, CA is lower (with marginal exceptions) across all values of *k* than for the baseline case where no clustering was used. It appears that the SVM is capable of accurately recognizing the complexities of within-genome compositional variations, and by clustering the data we are essentially reducing the relative number of training instances per class, thus negatively impacting SVM performance.

Table 4.4: Strict and Relaxed Classification Accuracies for Genome Pairs Processed Through the K-means Clustering Pipeline

For each genome pair, classification accuracies are presented for all k-means clustering SVM trials with k ranging from 2-6, as well as the baseline case for which no k-means clustering was performed (k=1). “Strict CA” indicates the classification accuracy given that the SVM was tasked to correctly classify each fragment to the correct cluster within the correct source genome. “Relaxed CA” shows the classification accuracy of the SVMs requiring only that each fragment be classified to the correct source genome regardless of cluster assignment.

Genome Pair	Baseline CA (k=1)	Strict CA					Relaxed CA				
		k=2	k=3	k=4	k=5	k=6	k=2	k=3	k=4	k=5	k=6
Methanosarcina barkeri str. Fusaro vs. Gramella forsetii KT0803	0.901	0.895	0.877	0.866	0.858	0.847	0.905	0.894	0.895	0.896	0.885
Ehrlichia ruminantium str. Welgevonden v2 vs. Methanosphaera stadtmanae DSM 3091	0.923	0.908	0.891	0.879	0.868	0.863	0.914	0.909	0.902	0.901	0.897
Buchnera aphidicola str. Cc (Cinara cedri) vs. Candidatus Sulcia muelleri GWSS	0.847	0.828	0.798	0.805	0.761	0.778	0.842	0.820	0.820	0.810	0.826
Prochlorococcus marinus str. MIT 9515 vs. Borrelia afzelii PKo	0.890	0.826	0.809	0.848	0.836	0.832	0.841	0.826	0.871	0.866	0.879
Chlamydophila abortus S26/3 vs. Neorickettsia sennetsu str. Miyayama	0.865	0.841	0.809	0.801	0.791	0.796	0.860	0.838	0.855	0.845	0.846
Prochlorococcus marinus str. AS9601 vs. Candidatus Pelagibacter ubique HTCC1062	0.859	0.843	0.819	0.795	0.798	0.778	0.854	0.840	0.837	0.835	0.816
Bradyrhizobium japonicum USDA 110 vs. Mesorhizobium loti MAFF303099	0.806	0.793	0.793	0.777	0.770	0.759	0.801	0.808	0.799	0.798	0.790
Lactobacillus acidophilus NCFM vs. Lactobacillus gasseri ATCC 33323	0.666	0.673	0.660	0.660	0.631	0.638	0.679	0.673	0.683	0.653	0.673
Haloarcula marismortui ATCC 43049 vs. Halobacterium salinarum R1	0.873	0.863	0.848	0.837	0.827	0.824	0.873	0.869	0.868	0.868	0.865
Prochlorococcus marinus str. MIT 9303 vs. Prochlorococcus marinus str. AS9601	0.975	0.958	0.947	0.938	0.931	0.912	0.972	0.972	0.974	0.973	0.971
Haemophilus somnus 2336 vs. Pediococcus pentosaceus ATCC 25745	0.931	0.916	0.899	0.887	0.874	0.855	0.927	0.928	0.929	0.925	0.920
Borrelia duttonii Ly vs. Borrelia recurrentis A1	0.559	0.520	0.500	0.544	0.523	0.523	0.533	0.505	0.559	0.529	0.547
Nitrobacter hamburgensis X14 vs. Nitrobacter winogradskyi Nb-255	0.688	0.688	0.658	0.646	0.623	0.617	0.678	0.675	0.665	0.643	0.645
Shewanella baltica OS195 vs. Shewanella denitrificans OS217	0.784	0.773	0.763	0.742	0.736	0.735	0.782	0.784	0.769	0.770	0.776
Nitrosospira multiformis ATCC 25196 vs. Nitrosomonas eutropha C91	0.893	0.870	0.870	0.843	0.853	0.844	0.886	0.889	0.886	0.901	0.898
Pyrococcus abyssi GE5 vs. Metallosphaera sedula DSM 5348	0.917	0.912	0.899	0.886	0.869	0.863	0.927	0.924	0.922	0.916	0.914

Despite the fact that clustering does not enhance the distinguishability of a given pair of genomes, confusion matrices from the clustered SVM trials show that fragments from certain clusters are preferentially misclassified into clusters in the respective comparator genomes (Figures 4.5-7). The misclassification of fragments from a given genome into clusters from the same genome (within-genome misclassification), tended to be much lower than the misclassification of fragments into clusters from the comparator genome (between-genome misclassification). For the three genome pairs examined in Figures 4.5-7, within-genome misclassification ranged from 2.67% - 4.70%, with *P. ubique* exhibiting the lowest rate of misclassification and *P. marinus* exhibiting the highest rate of within-genome misclassification, respectively. The relatively low rate of within-genome misclassification indicates that the clusters are well-formed in terms of the compositional features that define each cluster. Between-genome misclassification ranged from 7.7% (*H. marimortuii* fragments misclassified as *H. salinarum*) to 22.85% (*H. salinarum* fragments misclassified as *H. marismortui*).

Inspection of G+C content as well as the relative percentages of plus-strand and minus-strand coding bases (gene orientation bias) for each of the clusters indicates that fragments tend to be misclassified into clusters that have similar gene orientation biases and G+C content as the source cluster (Table 4.5). For example, the ribbons in Figure 4.5 show that fragments from cluster *a3* (*H. marismortui*) are preferentially misclassified into cluster *b1* (*H. salinarum*), and vice versa. Fragments from these two clusters are very similar in both the relative percentages of plus-strand and minus-strand coding bases (52% plus-strand, 48% minus strand for *H. marismortui;* 53.4% plus-strand, 46.6% minus-strand for *H. salinarum*). Additionally, the G+C content for fragments in these two clusters is quite similar (50.7% vs. 55.2%), and in both genomes these are the lowest observed G+C contents across all 6 clusters. Similar trends can be observed for misclassification between clusters *a4* and *b6*, as well as *a5* and *b2*. In both instances, fragments from the corresponding clusters exhibit extreme biases in gene orientation along with comparable G+C. Preferential cluster misclassification is also observed for the other two genome pairs: a1/b2, a1/b6, a2/b4, a4/b5, a5/b3, a6/b4 for *P. marinus* vs. *P. ubique* (Figure 4.6, Table 4.5), and a4/b3, a5/b6, a6/b2 for *E. ruminantium* vs. *M.*

Figure 4.5: Visualization of Cluster Misclassification for Haloarcula marismortui ATCC 43049 vs. Halobacterium salinarum R1.

This figure presents a visual representation of the confusion matrix for *H. marismortui* (clusters a1-a6) vs. *H. salinarum* (cluster b1-b6) for the k=6 trial. Clusters are arranged as arcs around the circumference of the figure. The length of a cluster represents the proportion of all genomic fragments assigned to that cluster during the k-means clustering step. Each colored ribbon represents the misclassification of fragments from one cluster to another, where the color of the ribbon denotes the true identity of the associated fragments, and the opposite end of each ribbon denotes the cluster assignment as predicted by the SVM. The width of a ribbon extending outward from a cluster with the same color indicates the overall proportion of fragments from the given cluster that were misclassified.

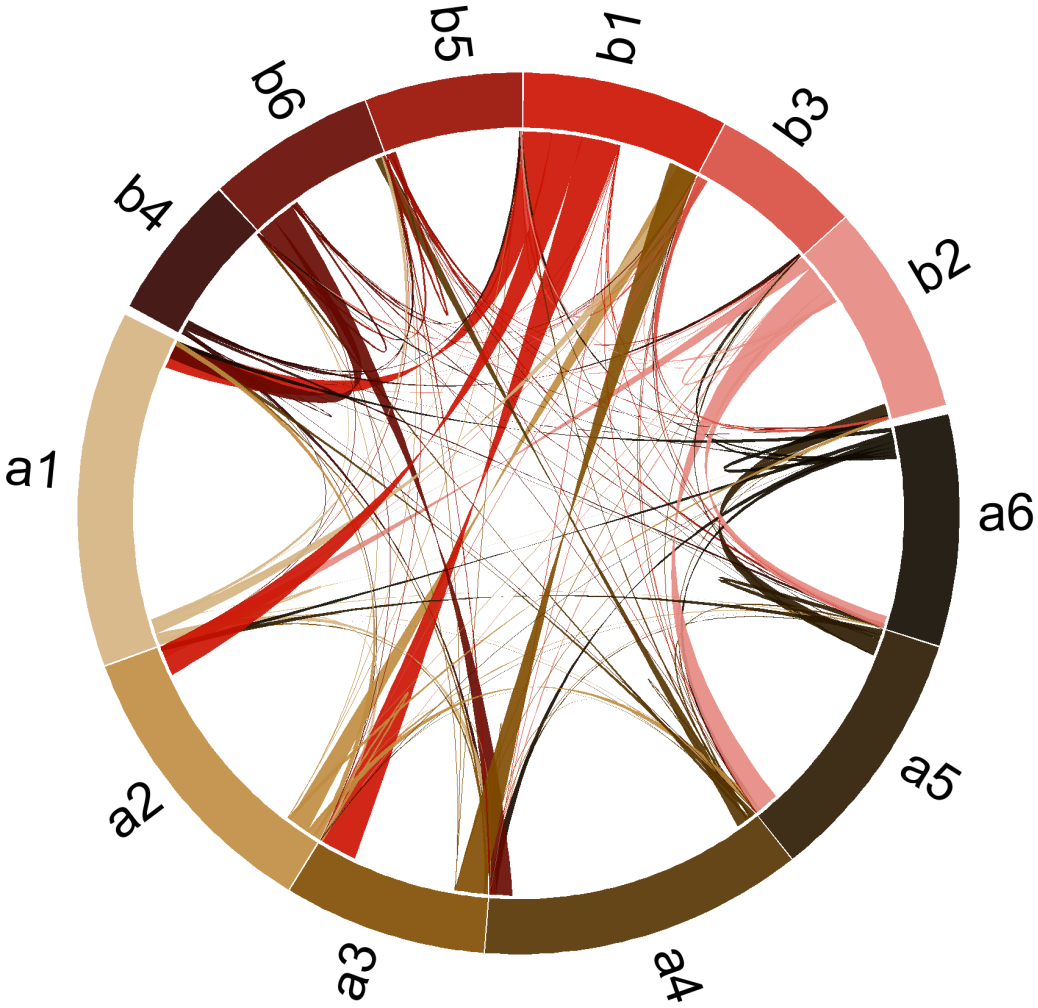


Figure 4.6: Visualization of Cluster Misclassification for Prochlorococcus marinus str. AS9601 vs. Candidatus Pelagibacter ubique HTCC1062.

This figure presents a visual representation of the confusion matrix for *P. marinus* (clusters a1-a6) vs. P. ubique (cluster b1-b6) for the k=6 trial. Clusters are arranged as arcs around the circumference of the figure. The length of a cluster represents the proportion of all genomic fragments assigned to that cluster during the k-means clustering step. Each colored ribbon represents the misclassification of fragments from one cluster to another, where the color of the ribbon denotes the true identity of the associated fragments, and the opposite end of each ribbon denotes the cluster assignment as predicted by the SVM. The width of a ribbon extending outward from a cluster with the same color indicates the overall proportion of fragments from the given cluster that were misclassified.

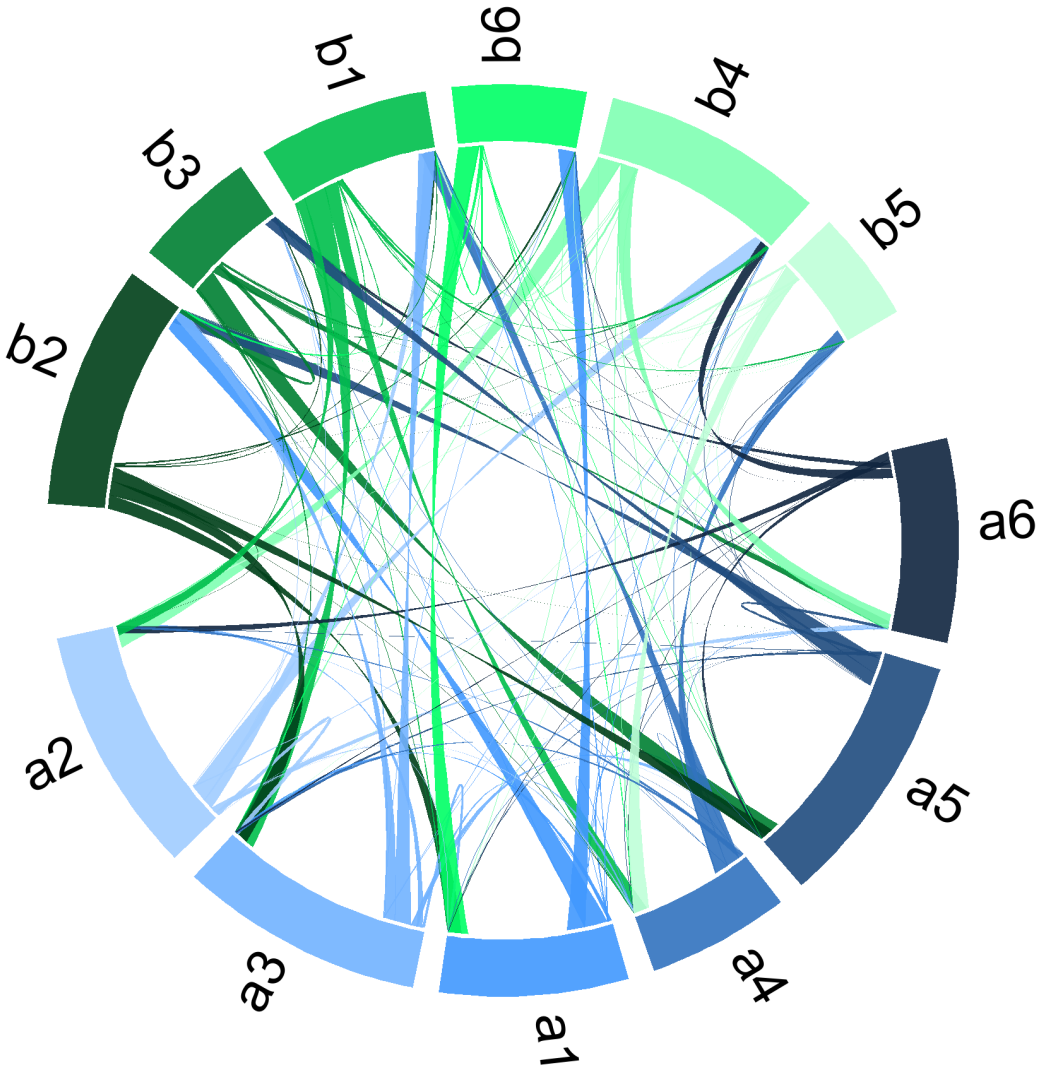


Figure 4.7: Visualization of Cluster Misclassification for Ehrlichia ruminantium str. Welgevonden v2 vs. Methanosphaera stadtmanae DSM 3091.

This figure presents a visual representation of the confusion matrix for *E. ruminantium* (clusters a1-a6) vs. *M. stadtmanae* (cluster b1-b6) for the k=6 trial. Clusters are arranged as arcs around the circumference of the figure. The length of a cluster represents the proportion of all genomic fragments assigned to that cluster during the k-means clustering step. Each colored ribbon represents the misclassification of fragments from one cluster to another, where the color of the ribbon denotes the true identity of the associated fragments, and the opposite end of each ribbon denotes the cluster assignment as predicted by the SVM. The width of a ribbon extending outward from a cluster with the same color indicates the overall proportion of fragments from the given cluster that were misclassified.

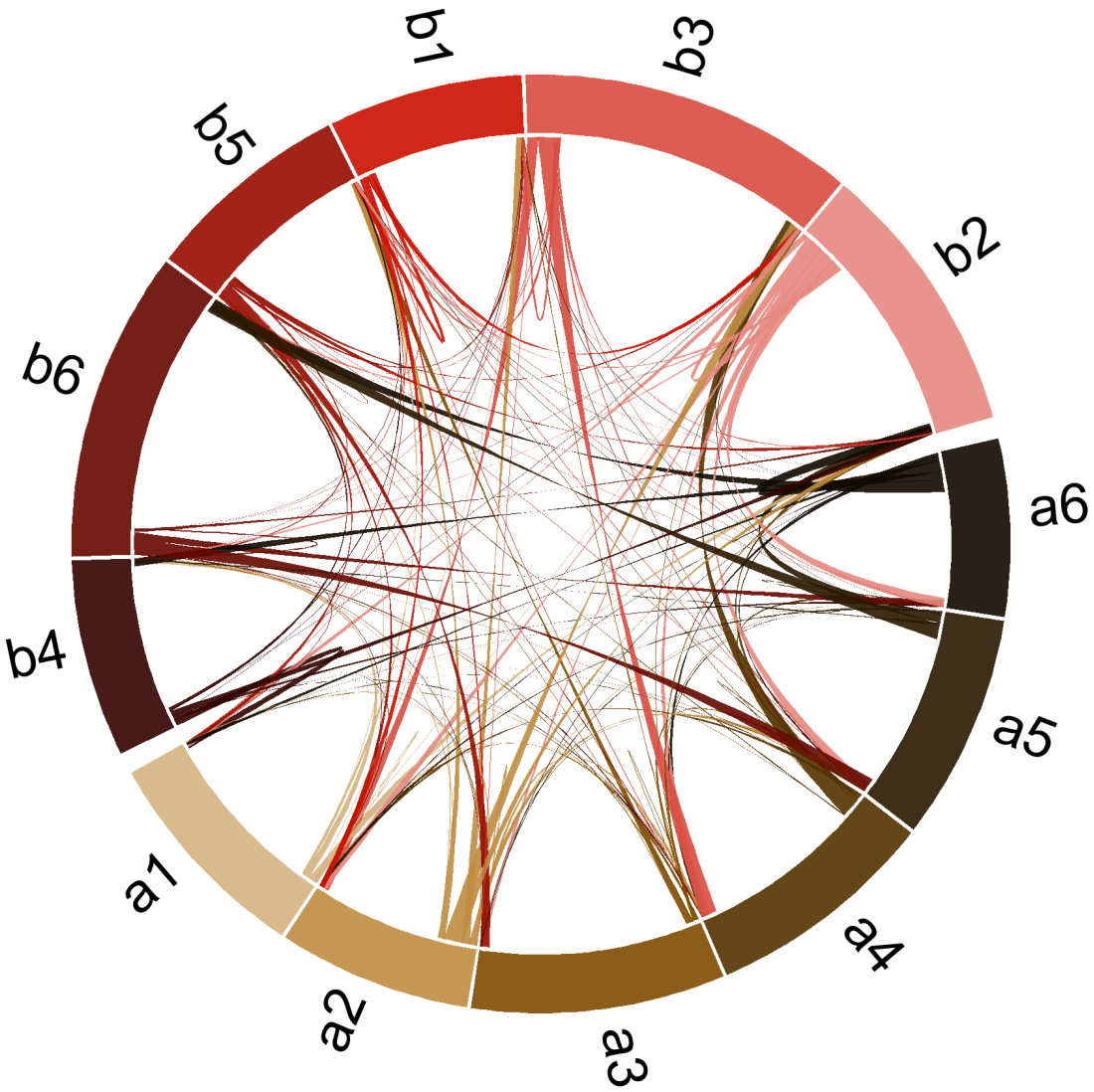


Table 4.5: Breakdown of Total Plus and Minus Strand Coding Nucleotides and %G+C Content by Cluster for H. marismortui vs. H. salinarum, P. marinus vs. P. ubique, and E. ruminantium vs. M. stadatmanae.

The number of coding nucleotides on the forward and minus strands are shown across all fragments belonging to each of the clusters from the k-means clustering analysis with k=6. For each cluster, 'nt' indicates the total number of coding nucleotides on each strand, while '%' denotes the percentage of the total coding nucleotides that exist on each strand. % G+C shows the average G+C content of all fragments assigned to a given cluster.

	H. marismortui ATCC 43049					*H. salinarum* R1				
Cluster	**+ Strand**		**- Strand**		**% G+C**	**+ Strand**		**- Strand**		**% G+C**
	nt	**%**	**nt**	**%**		**nt**	**%**	**nt**	**%**	
1	563330	68.9	254574	31.1	60.5	236701	53.4	206165	46.6	55.2
2	164830	26.0	468563	74.0	58.7	90938	17.4	430705	82.6	64.8
3	178620	52.0	164850	48.0	50.7	79138	21.3	291878	78.7	70.9
4	704869	88.5	91870	11.5	65.1	268159	74.3	92881	25.7	70.8
5	33037	5.2	602104	94.8	63.4	176374	45.7	209944	54.3	71.2
6	225505	41.2	322063	58.8	65.8	366430	87.7	51184	12.3	65.6

	P. marinus AS9601					*P. ubique HTCC1062*				
Cluster	**+ Strand**		**- Strand**		**% G+C**	**+ Strand**		**- Strand**		**% G+C**
	nt	**%**	**nt**	**%**		**nt**	**%**	**nt**	**%**	
1	199295	82.7	41789	17.3	27.4	91616	42.8	122307	57.2	27.8
2	37726	11.8	282772	88.2	31.6	280861	88.7	35931	11.3	30.4
3	196543	63.6	112636	36.4	29.3	127705	80.0	31848	20.0	35.3
4	35971	20.3	140890	79.7	25.6	6327	2.2	277366	97.8	32.4
5	306286	92.0	26722	8.0	35.1	15900	11.0	128890	89.0	25.2
6	35687	13.6	225904	86.4	36.3	148147	86.4	23262	13.6	24.8

	E. ruminantium Welgevonden v2					*M. stadtmanae* DSM 9091				
Cluster	**+ Strand**		**- Strand**		**% G+C**	**+ Strand**		**- Strand**		**% G+C**
	nt	**%**	**nt**	**%**		**nt**	**%**	**nt**	**%**	
1	32502	52.2	29754	47.8	22.5	192605	94.5	11238	5.5	26.6
2	134940	69.5	59136	30.5	28.6	85254	29.5	204154	70.5	27.3
3	18450	31.5	40120	68.5	22.7	369056	96.3	14125	3.7	30.9
4	240174	95.0	12693	5.0	31.6	19856	11.3	156521	88.7	23.5
5	11302	4.6	233486	95.4	31.4	102257	71.5	40852	28.5	21.7
6	49006	26.9	133330	73.1	28.8	6164	1.7	349382	98.3	31.7

stadtmanae (Figure 4.7, Table 4.5).

In certain cases, clusters involved in preferential misclassification showed similarities in G+C biases while differing greatly in the terms of the gene-orientation biases, such as in the case of *b1* fragments being misclassified into clusters *a1* and *a2* in Figure 4.5. Although these clusters differ substantially in terms of their relative percentage of plus/minus strand coding bases (Table 4.5), the clusters contain relatively low G+C in relation to the other clusters. Similarly, in Figure 4.7, fragments from preferentially misclassified clusters *a3* and *b5* have very different gene orientation biases (31.5% plus-strand, 68.5%% minus-strand; 71.5% plus-strand, 28.5% minus-strand) although in this case both clusters show extremely low G+C contents (22.7%; 21.7%).

Distribution of Correctly Versus Incorrectly Classified Fragments Within a Genome

In the comparison between *H. salinarum* and *H. marismortui*, regions of misclassification are distributed non-uniformly across both genomes (Figure 4.8). The majority of the regions appear within the *H. salinarum* genome, particularly on all four plasmids as well as the region between nucleotides 15,000-70,000 on the primary chromosome. The density of misclassified fragments on the *H. salinarum* plasmids suggests that the plasmids are closer in composition to *H. marismortui* than to the actual source genome. This may reflect a recent transfer of genetic material from *H. marismortui* to H. *salinarum*. Additionally, the presence of a large region of misclassification near the start of the main chromosome in *H. salinarum* may in fact represent the integration of part of one of the plasmids. Given sufficient time, DNA acquired from another microbe will eventually ameliorate and become indistinguishable from the rest of the genome [55]. The fact that the highlighted regions within *H. salinarum* are so predominantly localized supports the idea that a recent transfer may be responsible for the misclassified fragments. Similar plots for *Ehrlichia ruminantium* str. Welgevonden v2 vs. *Methanosphaera stadtmanae* DSM 3091 (Figure 4.9) and *Prochlorococcus marinus* AS9601 (13548) vs. *Pelagibacter ubique* (Figure 4.10) demonstrate much more uniform distributions of misclassified fragments across each

Figure 4.8: Correct and Incorrect Classifications Versus Genome Position for 500 nt Fragments From Haloarcula marismortui and Halobacterium salinarum.

In this figure, arcs are used to represent the genomes involved in the underlying SVM pairwise comparison. Arcs with a light grey border on the exterior face represent sequences from *H. salinarum*, whereas a dark grey border indicates sequences from *H. marismortui*. Along the length of each genome, green segments indicate correct classifications while red segments indicate misclassified fragments. “Regions of misclassification” are indicated by blue bars within the interior of the figure, and represent spans of each genome that contain > 37.5% misclassified fragments.

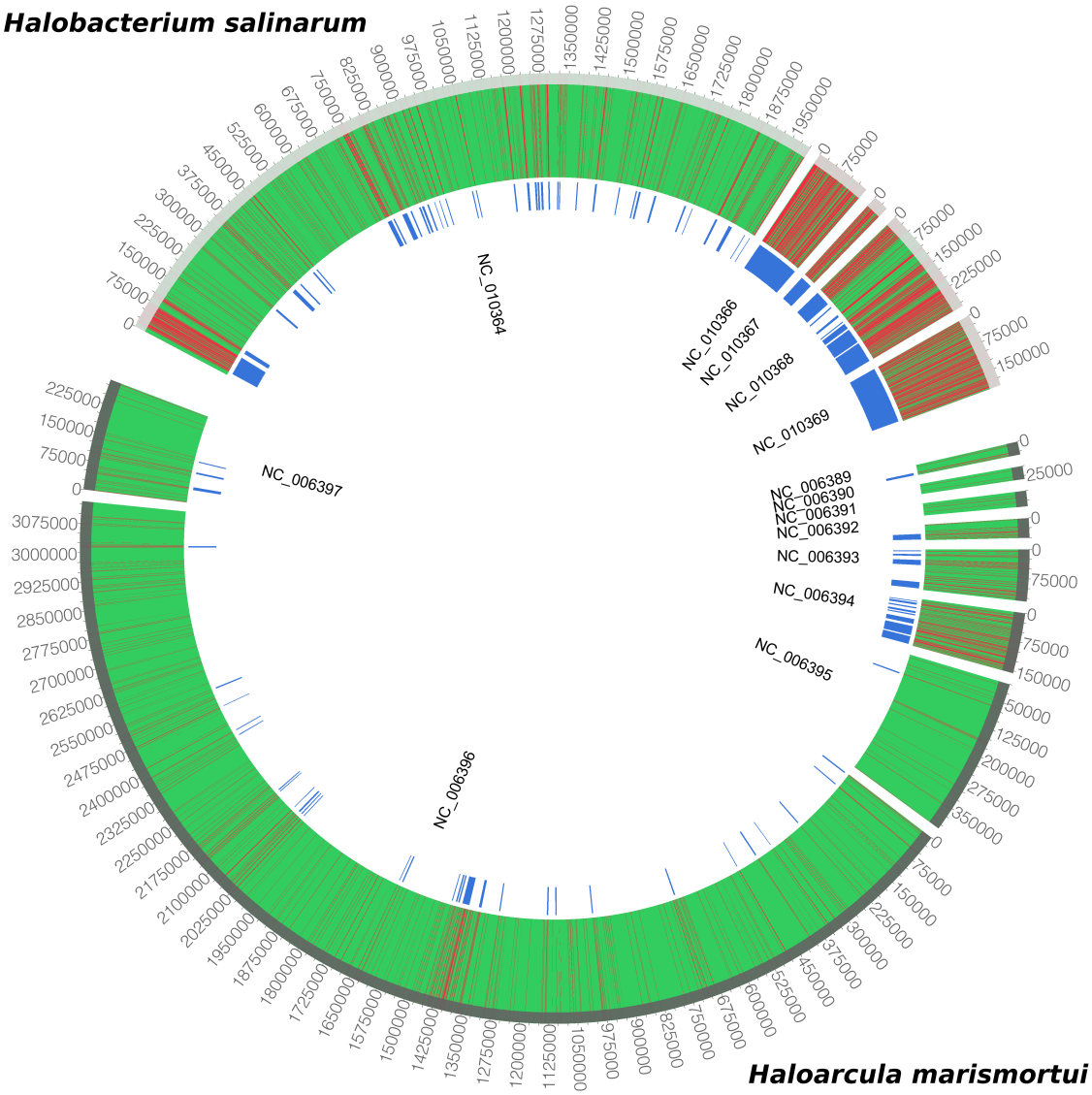


Figure 4.9: Correct and Incorrect Classifications Versus Genome Position for 500 nt Fragments From M. stadtmanae (15579) and E. ruminantium str. Welgevonden (13355)

In this figure, arcs are used to represent the genomes involved in the underlying SVM pairwise comparison. Along the length of each genome, green segments indicate correct classifications while red segments indicate misclassified fragments. “Regions of misclassification” are indicated by blue bars within the interior of the figure, and represent spans of each genome that contain > 37.5% misclassified fragments.

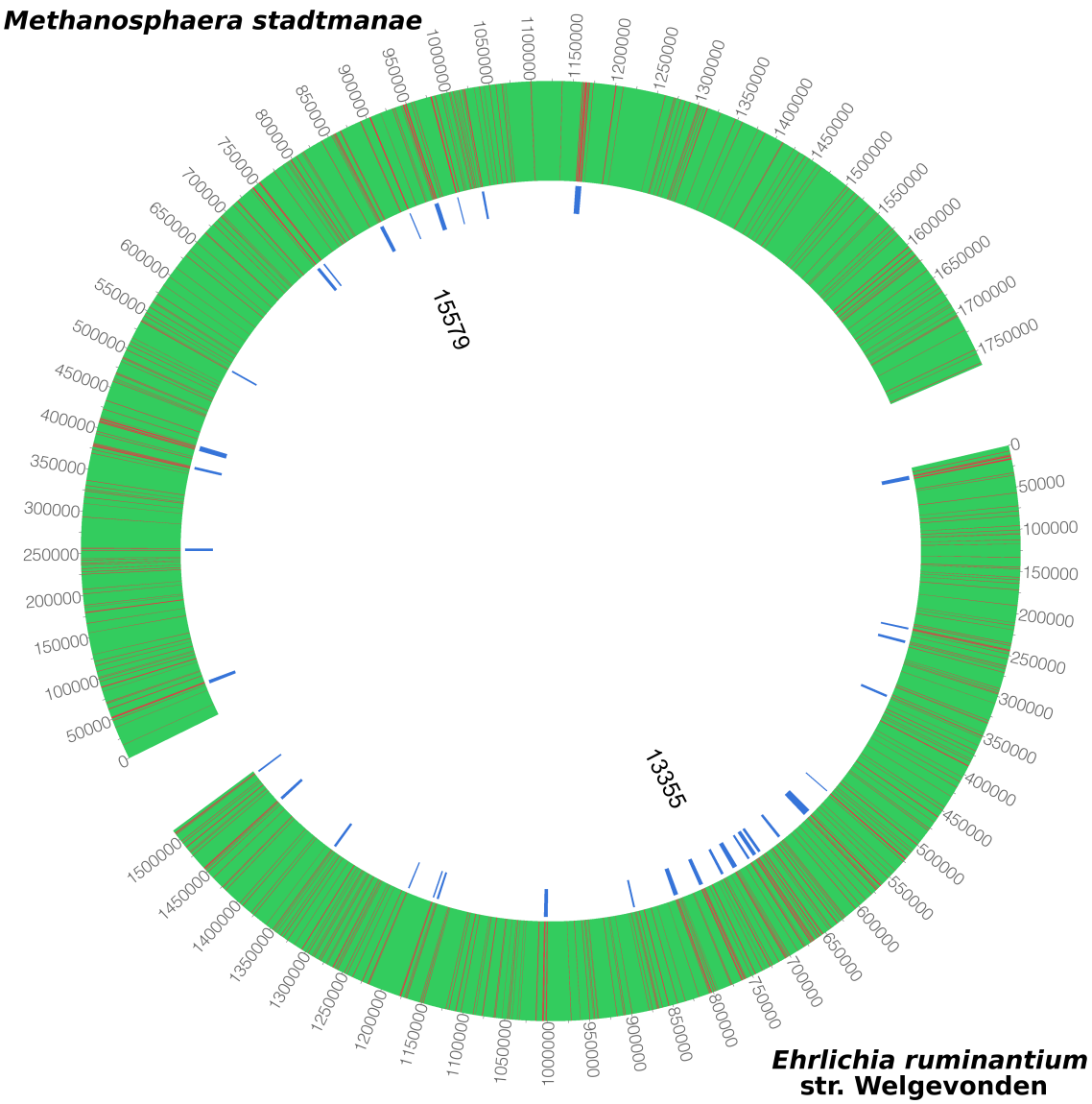
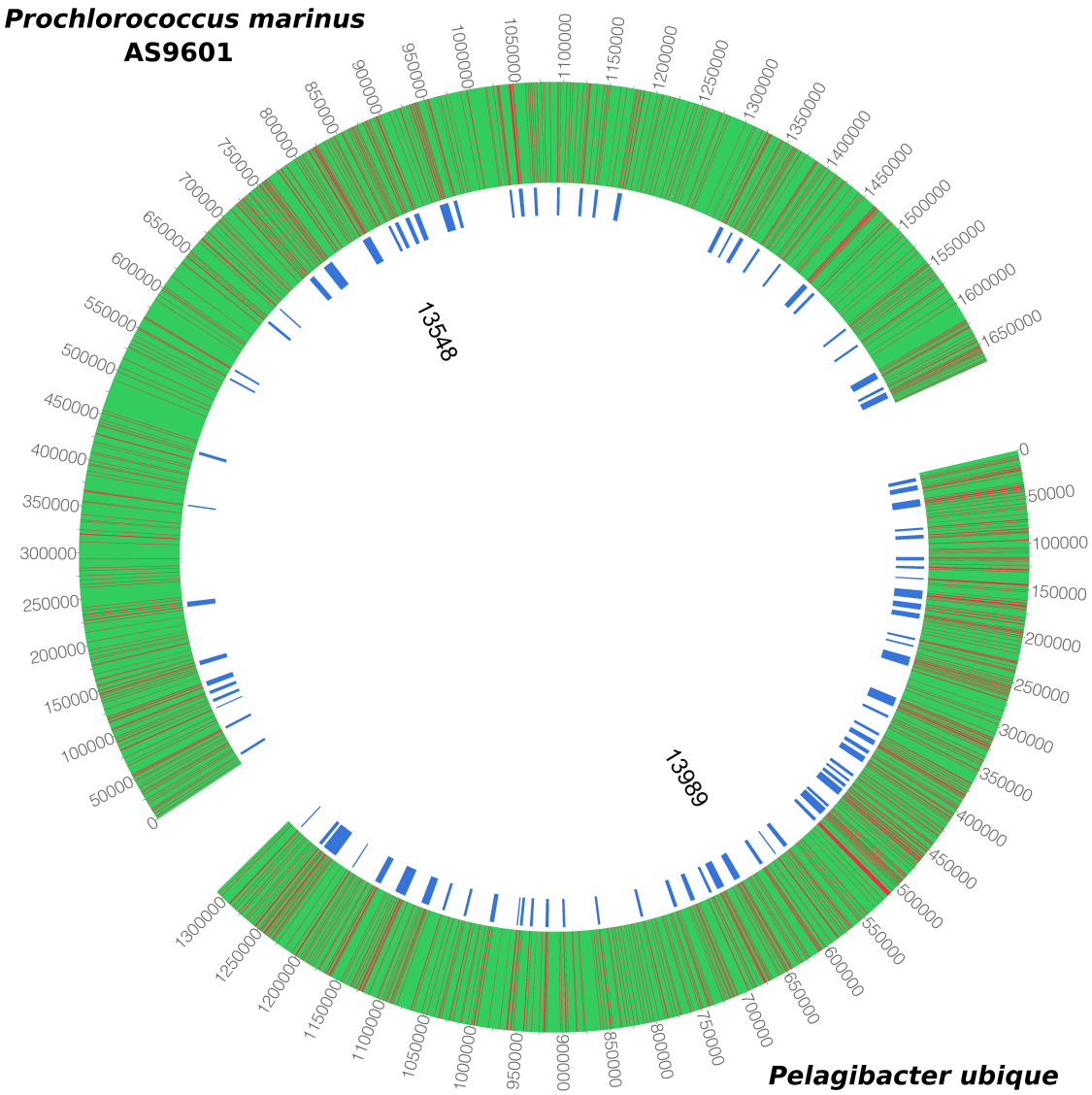


Figure 4.10: Correct and Incorrect Classifications Versus Genome Position for 500 nt Fragments From P. marinus AS9601 (13548) and P. ubique (13989)

In this figure, arcs are used to represent the genomes involved in the underlying SVM pairwise comparison. Along the length of each genome, green segments indicate correct classifications while red segments indicate misclassified fragments. “Regions of misclassification” are indicated by blue bars within the interior of the figure, and represent spans of each genome that contain > 37.5% misclassified fragments.



genome, and do not support the notion of recent clustered LGT events for these pairs of genomes.

nBLASTP Score Distributions for Orthologs That Fall Within Regions of Misclassification

For *H. salinarum* vs. *H. marismortui*, the distribution of nBLASTP scores for orthologs overlapping regions of misclassification differs from that of the nBLASTP scores for orthologs that do not overlap such regions (Figure 4.11a). For correctly classified fragments, the histogram and density distribution both show a single peak at 0.65. Incorrect fragments, on the other hand, show peaks at both 0.4 and 0.75, suggesting that the incorrectly classified orthologs belong to two groups: a set of more distantly related orthologs with lower nBLASTP scores, and a more closely related set of orthologs with higher nBLASTP scores. This group of closely related orthologs may represent genes involved in a recent LGT event which have not yet undergone sufficient amerlioration to bring their compositions in line with the acceptor genome. Neither of the other genome pairs (Figure 4.11b,c) shows this two-peak distribution for incorrectly classified fragments. Mann-Whitney tests comparing the nBLASTP distributions between correctly and incorrectly classified fragments (Table 4.6) indicate that only the distributions for *H. salinarum* and *H. marismortui* are statistically different ($p = 5.348 \times 10^{-9}$).

Figure 4.11: Distributions of Reciprocal Best Hit nBLASTP Scores for Putative Orthologs

For each pair of genomes, histograms and population density plots were generated for the nBLASTP scores of all orthologs, orthologs that were correctly classified, and orthologs that were incorrectly classified. Orthologs that have at least 95% nucleotide overlap with a region of misclassification were deemed 'incorrect' whereas all remaining orthologs were deemed 'correct'. a) *Halarcula marismortui* ATCC 43049 vs. *Halobacterium salinarum* R1, b) *Ehrlichia ruminantium* str. Welgevonden v2 vs. *Methanosphaera stadtmanae* DSM 3091, c) *Prochlorococcus marinus* str. AS9601 vs. *Candidatus Pelagibacter ubique* HTCC1062

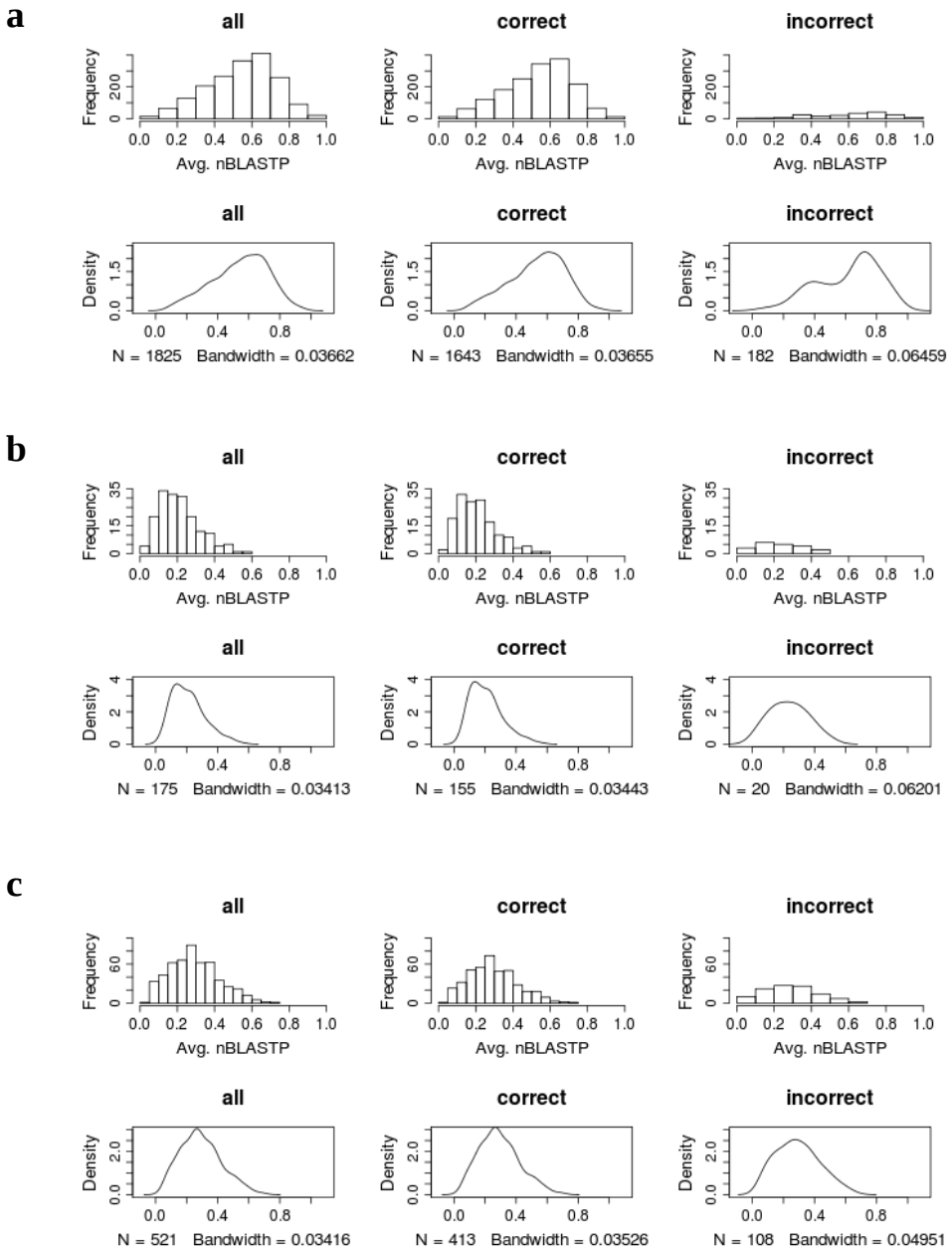


Table 4.6: Results of 2-sided Mann-Whitney Tests Comparing the Distributions of nBLASTP Scores for Correctly Versus Incorrectly Classified Fragments

Genome Pair	# correctly classified orthologs	# incorrectly classified orthologs	p-value
H. marismortui vs. *H. salinarum*	1643	182	5.348e-09
M. stadtmanae vs. *E. ruminantium*	155	20	0.2735
P. marinus vs. *P. ubique*	413	108	0.8816

Conclusions

Despite the fact that the pairwise comparisons in this set of experiments were based upon tetranucleotide frequency profiles from relatively short 500-nt genomic fragments, most genome pairs achieved excellent distinguishability, with 93.2% of the 299,151 SVM runs resulting in a classification accuracy of 95% or greater. This is not surprising, as the majority of comparisons were performed between pairs of distantly related organisms that have had ample opportunity to diverge in terms of their composition. Likewise, the mean CA of 98.1% observed across all comparisons is encouraging, but it mainly reflects this overwhelming majority of comparisons between distantly related genomes, which are generally the most trivial cases for the SVM classifier. Comparisons between congeners, on the other hand, typically lead to very poor distinguishability, except for genome pairs that have undergone rapid divergence, such as strains of *Prochlorococcus marinus* or *Clostridium botulinum*. These trends are consistent with previous studies that showed significant correlation between the similarity of phylogenetically relevant marker genes (such as 16S) and various measures of compositional similarity [130; 131]. Under certain circumstances, distantly related organisms may converge in their genome composition, due to factors such as extreme G+C or A+T content, crowding of the oligonucleotide space [80], and habitat convergence [132], however this convergence does not appear to interfere with classification in the majority of cases.

In agreement with previous studies [36; 38; 62], the results in this chapter demonstrate a trend of decreasing classification accuracy with increasing convergence of genome signature. When CA is examined in terms of the various measures of genomic similarity considered in this study (G+C content, 16S distance, tetramer distance, and lowest common taxonomic rank), all support this general trend. Although the CA vs. tetramer Euclidean distance model had a slightly higher R^2 value than that of the CA vs. 16S distance model, the 16S model was ultimately used in order to select interesting outliers, due to the high level of phylogenetic signal carried within 16S rDNA sequences [28]. Tetramer frequencies, although powerful in their ability to distinguish between

fragments based on composition alone, have been shown to carry very little phylogenetic signal [40]. The difference in R^2 values is likely a result of differences in the scope of genome signature captured by these two methods. 16S distance is calculated using a small number (normally 2) of highly-conserved marker genes which makes it useful for resolving taxonomic relatedness, but is unlikely to reflect the true compositional variation observed throughout the entire genome. Tetramer Euclidean distance, on the other hand, is calculated from the average tetramer distances observed across each genome in a given comparison, and is more likely to represent the global compositional patterns of each genome than l6S distance, while sacrificing the taxonomic specificity offered by the marker gene approach.

When the set of correctly classified fragments was interpreted in terms of the functional categories of their encoded proteins, fragments encoding genes involved in biosynthetic pathways (e.g., amino acid biosynthesis, synthesis of cofactors) tended to be overrepresented in this group. The composition of genes in this category is much less constrained in comparison to that of informational genes, and this relative lack of compositional constraint facilitates an organism's ability to adapt its biochemical pathways in order to adjust to new energy sources or to develop antibiotic resistance, for example. Misclassified fragments showed an overrepresentation of functional categories that are likely to differ from the core genome signature for each organism, such as informational genes involved in core cellular processes or signal transduction, and proteins associated with mobile and extrachromosomal elements. In the case of informational genes, proteins within this class are known to evolve very slowly [133], and are highly constrained by interactions with other core proteins within the cell. Although synonymous mutations may provide some baseline level of compositional divergence, informational genes are still unlikely to exhibit the core genome signature, which in turn is likely to reduce the classification accuracy of such fragments. Similarly, mobile elements and other introgressed sequence are likely to have compositions that differ greatly from the host genome. If such sequences are found in both genomes from a given genome pair, the foreign sequences are likely to be classified into one genome or another in an arbitrary fashion, thus reducing the overall classification accuracy of

fragments encoding genes within these classes. Fragments that contain mixtures of coding and noncoding sequence tend to be more difficult to classify than fragments comprising coding sequence from a single open reading frame. Correctly classified fragments from very closely related organisms showed an increased tendency for fragment heterogeneity in some instances, suggesting that noncoding sequences might have an important influence on classification for these exceptional cases.

This study clearly demonstrated that the unsupervised compositional clustering of genomic fragments prior to SVM classification offers no increase in classification accuracy, indicating that the SVM is sufficiently powerful to model the complete set of compositional classes that are produced by the unsupervised clustering step. Although clustering does not improve classification accuracy, the examination of the cluster confusion matrices brings light to a number of interesting characteristics of the tetranucleotide frequency data. In most cases, when a fragment is classified into the wrong genome, the fragment is preferentially assigned into a compositional cluster that is similar in terms of both the gene strand orientation bias and G+C content of the true source cluster. Additionally, the low rate of within-genome misclassification during the clustering experiment indicates that the compositional clusters identified during the k-means clustering step are in fact well defined in terms of their compositional characteristics, and do not simply represent random associations of fragments into one of the 6 possible clusters. In cases where misclassified fragments belong to clusters with extremely high or low G+C, these fragments tend to be misclassified into similarly biased clusters in the comparator genome without regard for gene orientation.

'Symmetrization' of tetranucleotide frequencies and correction for fragment level G+C content offered little or no increase in classification accuracy in the present study, despite the fact that such techniques have been implemented by several pre-existing methods [37; 38; 63]. The SVM appears to be sufficiently robust to capture the underlying genome signature without regard to strand biases or local G+C content, a characteristic that may not be shared with less complex classifiers such as TETRA [40]. Classification accuracy improved in the case where tetranucleotide frequencies were

adjusted using the genomic G+C content, however this observation is unlikely to be useful in developing an improved classifier, as the genomic G+C content will not usually be known when attempting to classify anonymous DNA fragments.

The results from this chapter suggest that in specific cases, the ability to distinguish between fragments from distinct genomes may be fundamentally limited due to a lack of compositional divergence, or the effect of various constraints imposed on genome signature. Examples of such fragments that are likely to be misclassified include the fragments that overlap highly conserved informational genes, fragments from genomes with extremely high or low G+C, or fragments containing sequence that has not yet undergone sufficient amelioration. Although it is unlikely that the classification of such fragments will improve substantially using improved classifiers, knowledge of such difficult cases may be useful in attempting to improve the classification accuracy of boundary cases represented by fragments that contain mixtures of one or more compositional classes of genome sequence. The results from the fragment heterogeneity experiment suggest that classification accuracy might be improved if the underlying fragments were partitioned in order to contain sequence belonging to a single compositional class rather than mixtures of one or more classes. Sequence segmentation has been implemented previously for other applications [134] and will likely prove useful in the classification of DNA fragments. An improved classifier might use various preprocessing steps in order to improve the homogeneity of DNA fragments prior to classification. In the case of coding vs. non-coding sequence, methods such as Orphelia [135] are able to identify microbial open reading frames with a high level of accuracy, and would be useful in partitioning DNA fragments into coding and non-coding bins. Fragments that overlap multiple open reading frames that differ in orientation may also show reduced classification accuracy due to corruption of the underlying genome signature. Given the well-defined clusters observed in the k-means clustering study, a novel algorithm might examine the sequence of each fragment in order to partition the fragment into subsequences that represent plus-strand and minus-strand encoded ORFs, thus further reducing the heterogeneity of the tetranucleotide signature and improving distinguishability. As indicated by the functional annotation analysis, certain functional

classes of genes are inherently difficult to classify, such that the classification accuracy of fragments containing mixtures of easily classified and difficult to classify fragments may be impeded by the presence of the more compositionally constrained functional classes. Homology searches using the BLASTP [71; 72] or PFAM [74] databases might help to identify these classes within metagenomic fragments, allowing the homogeneity of these fragments to be improved by partitioning the fragments into the subsequences representing the component classes. Furthermore, if one of the subsequences is determined to belong to an easily classified functional category, the assignment of such a sequence might reasonably be projected onto a less easily classified subsequence derived from the same genomic fragment, thus improving classification accuracy.

Chapter 5 – Discussion

Summary of Experiments

This thesis presents three experiments designed to identify factors that influence the relative distinguishability of microbial genomes based on patterns of genomic composition. All experiments made use of the support vector machine, a supervised, state-of-the-art machine learning method that has successfully been applied to a wide variety of classification problems [96-99]. The SVM is particularly well-suited to composition-based classification due to its ability to generate robust models for relatively large and complex feature sets, such as the k-mer frequency profiles used in Chapters 2-4. Notably, the SVM also serves as the underlying classification strategy for PhyloPythia [38], the most accurate metagenomic DNA classification system to date.

Multi-class classifiers are useful in situations where we would like to measure the overall performance of a classifier in response to various parameter changes, without focusing specifically on the underlying pairwise comparisons. Such is the case for the experiment outlined in Chapter 2, where a multi-class SVM was used to evaluate the impact of DNA recoding schemes, fragment length, and k-mer length on the global classification accuracy for a set of 10 microbial genomes. In contrast, the aim of Chapters 3 and 4 was to identify specific pairs of genomes that demonstrate higher or lower classification accuracy than might be predicted by models that relate classification accuracy with various measures of genomic similarity. In these cases, pairwise SVMs were required in order to examine classification accuracy for each possible pairing of genomes, details that are not readily available when using a multi-class SVM. For a given multi-class data set, pairwise classifiers will result in higher classification accuracies than that of a single a multi-class classifier, and will also provide an indication as to which specific pairs of classes are easier or more difficult to distinguish relative to the complete set of comparisons.

For any given pair of genomes, the DNA sequence of each genome can be divided into one of two general classes: 1) orthologous sequences that have been inherited from

the most-recent common ancestor and thus are represented in both genomes; and 2) non-orthologous sequences that are unique to a single genome in the pair, which may be the result of such factors as phage integration, LGT, genomic rearrangement, or deletion. When comparing the relative distinguishability of a pair of genomes, it is useful to consider classification accuracy based on the set of core orthologous sequences, removing the influence of non-orthologous sequences on the given classifier. Such is the case for the experiment described in Chapter 3, where classification was based on the tetranucleotide frequency profiles for the orthologous sequences shared by each pair of genomes. The experiment in Chapter 4 used an alternative approach, and examined the classification accuracy as determined from the tetranucleotide profiles for genomic fragments, considering the combined influence of orthologous and non-orthologous sequences on pairwise classification. Such raw genomic fragments are comparable in length to the sequencing reads generated by high-throughput metagenomic sequencing projects.

Pre-processing of feature sets prior to classification is a common practice whereby the data sets to be classified can be modified in order to remove certain biases in the data or to otherwise increase the suitability of the data for classification. Common approaches to pre-processing include scaling, clustering, recoding, and normalization. The experiment in Chapter 2 used various recoding strategies in order to determine whether the use of degenerate k-mer patterns and increased k-mer length improves the classification accuracy of the multi-class SVM. In Chapter 4, unsupervised clustering of the tetranucleotide frequency profiles, 'symmetrization' of the k-mer frequencies, and correction for both fragment-level and genome-level G+C content were employed in order to determine their influence on the pairwise distinguishability of microbial genomes. In all cases except the correction for genome-level G+C content, any information required for the various pre-processing steps could be extracted directly from the data sets, without requiring properties derived from the complete genome sequences that gave rise to the DNA fragments. The lack of dependence of the pre-processing methods on the availability of complete genome sequences makes such methods appropriate for applications involving anonymous metagenomic DNA fragments.

Summary of Results

Chapter 2

The DNA recoding experiment in Chapter 2 demonstrated that multi-class SVM classifiers trained using various binary recoding schemes were able to distinguish between 10 bacterial genomes with high classification accuracy for long (5000 nt) fragments, however, performance was poor for short fragment lengths typical of metagenomic sequencing projects. Furthermore, the reduction in feature space and increase in usable pattern lengths provided by the DNA recoding techniques offered no increase in performance relative to the k-mer classifier, which outperformed the binary recoding based classifiers across the full range of fragment and pattern lengths considered in this experiment (with few exceptions). Consistent with previously reported findings [38], the results from this experiment indicated that composition-based SVM classifiers trained using oligonucleotide frequency profiles from short fragments were able to classify longer fragments with little decrease in accuracy, whereas classifiers trained using longer fragments performed poorly when faced with shorter fragments.

Chapter 3

The results from the pairwise comparisons in Chapter 3 demonstrated that in general, the distinguishability of a pair of microbial genomes based on the tetranucleotide profiles of orthologous sequences was proportional to 16S rDNA distance, G+C distance, and tetramer Euclidean distance, and inversely proportional to both the lowest common taxonomic rank and average nBLASTP scores. Analysis of outliers from the CA vs. average nBLASTP model identified a number of factors that may lead to an increase or decrease in distinguishability for a given genome pair, including similarities in habitat and lifestyle, tendency for genome rearrangement, lack of DNA repair enzymes, the reduced nature of obligate intracellular pathogens, unusual selective pressures or evolutionary strategies, extreme G+C content, and the presence of numerous repeats, truncated genes, or phage DNA in one or both of the genomes. Of the 4 pairs of outliers considered in the study, the two genome pairs that demonstrated less-than-expected

distinguishability comprised obligate intracellular pathogens that reside in vacuoles within a mammalian host cell. The remaining two outlier pairs showed higher than expected classification accuracies, and represented a pair of closely related, free-living aquatic bacteria which have adopted distinct evolutionary strategies, and a pair of intracellular pathogens that live freely within the cytosol of the host cell. The results from this chapter also indicate that distinguishability does not always correlate with marker gene-based measures of genomic similarity, such that compositional convergence or divergence caused by factors both intrinsic or external to the genome can have a significant impact on distinguishability.

Chapter 4

The experiment described in Chapter 4 demonstrated that a composition-based SVM classifier was capable of distinguishing between the vast majority of genome pairs with high accuracy, despite the fact that SVM models were trained using tetranucleotide frequency profiles for very short (500 nt) fragments. Conspecific comparisons generally resulted in poor classification accuracy, except in cases where the genomes had undergone rapid compositional divergence, as in the case of strains of *Prochlorococcus marinus* (*P. marinus* str. MIT 9303 vs. *P. marinus* AS9601; CA = 97.4%). As reported in Chapter 3, distinguishability was generally proportional to G+C distance, 16S rDNA distance, tetramer Euclidean distance, and inversely proportional to lowest common taxonomic rank. Although unsupervised clustering of tetranucleotide frequency profiles did not improve distinguishability, analysis of the resulting confusion matrices indicated that both the G+C content and the polarity of protein-coding sequences within a fragment can contribute to misclassification. The results from this chapter confirmed that fragments containing protein-coding sequence from certain functional role categories showed significant trends for correct or incorrect classification. Examination of fragment heterogeneity in relation to classification indicated that fragments containing multiple compositional signatures showed an increased tendency for misclassification. Exceptions to this trend were observed for pairs of very similar genomes, for which an increased proportion of faster-evolving non-coding sequence in the associated fragments may have

led to increased distinguishability and fragment heterogeneity. For a pair of genomes that demonstrated less-than-expected classification accuracy, an unusual distribution of average nBLASTP scores for misclassified sequences as well as a local clustering of misclassified fragments within one of the genomes supported the notion that a recent LGT event may have contributed to the observed decrease in distinguishability relative to the model. Symmetrization of oligonucleotide frequency profiles, a common practice used by several existing DNA classifiers, was shown to have little effect on classification accuracy. Correction of oligonucleotide frequency profiles based on fragment G+C content showed no change in performance, whereas an increase in distinguishability was noted for several genome pairs when the frequencies were corrected using genomic G+C.

Applications of Key Findings and Future Work

Existing composition-based DNA classification methods are likely to benefit from the results presented in this thesis. For the typical classifier, the first step in the classification process typically involves the construction of the training set and the parameterization of the corresponding DNA fragments into a form that is applicable to the underlying machine learning method. Many of the results reported here can be applied to this preliminary stage of classification in order to maximize distinguishability. For instance, results from Chapter 2 indicated that the use of recoding and degenerate k-mer patterns should be avoided, as both have been shown to decrease classification accuracy. Similarly, the unsupervised clustering of oligonucleotide frequency profiles prior to classification by advanced methods such as the SVM is likely to degrade performance when compared to the use of unclustered data. Clustering may be advantageous for less advanced classifiers such as TETRA [40] or the modified k-NN approach [37], as the performance of these methods may be impeded by the presence of distinct compositional bins within each genome. Despite the frequent use of techniques such as the symmetrization of oligonucleotide frequencies and correction for G+C content in the literature [38; 80; 128], the results presented in Chapter 4 showed that neither of these pre-processing steps offers an increase in classification accuracy over the unsymmetrized and uncorrected oligonucleotide frequency profiles when the SVM is

used as the underlying classifier.

For data sets consisting of fragments of various lengths, the results from Chapter 2 confirmed previously reported findings that it is advantageous to train models using the frequency profiles for shorter fragments, as the resulting models will be more generalizable to the classification of fragments of longer lengths [38]. Conversely, the use of longer fragments when training a classifier should be avoided, as the resulting models will show greatly reduced ability to accurately classify fragments shorter than those used to train the model. At present, composition-based classifiers calculate k-mer frequency profiles for DNA fragments without considering that the fragments may contain mixtures of non-coding sequence as well as coding sequence from one or more genes [36-38; 40; 52; 65]. As demonstrated in Chapter 4, fragment heterogeneity is associated with an increased tendency for misclassification, and as such, the use of a sequence-segmentation approach prior to the parameterization of DNA fragments into k-mer frequency profiles is likely to increase distinguishability by decreasing the heterogeneity of fragments containing more than one class of sequence. Additionally, the classification of fragments that contain protein-coding sequences from multiple genes in opposite orientations or genes that are associated with functional role categories that have a tendency for misclassification may also be improved using this sequence-segmentation approach.

Once a composition-based classifier has produced a set of predictions for a given data set, multiple characteristics identified in this thesis may be used to express an overall confidence in each of the predictions. For example, Chapter 4 demonstrated that fragments containing protein-coding sequence associated with certain biological functions showed a tendency for increased or decreased classification. Existing classifiers might be extended to use BLASTP [71] or PFAM-based [75] searches in order to identify DNA fragments that are associated with these biological roles, and assign an increase or decrease in confidence to the associated fragments. Likewise, Chapter 4 also showed that fragments exhibiting extreme G+C biases were more likely to be misclassified, and existing classifiers could be modified to report decreased confidence in such instances.

An important result reported in both Chapters 3 and 4 indicated that although

genome distinguishability could be modelled in relation to various measures of compositional similarity and taxonomic relatedness, the models were imperfect, and outliers were identified that had either increased or decreased compositional similarity than that suggested by their taxonomic relatedness. An improved classifier might combine both the semi-supervised approaches of CompostBin [65] and S-GSOM [36] with the supervised approach presented in PhyloPythia [38] to take both taxonomic and compositional similarities into account when classifying metagenomic fragments. For example, if multiple forms of a conserved marker gene are found within a set of fragments that show very high similarity in patterns of genomic composition, the assignment of such fragments could be augmented with information regarding the number of likely genomes (and their taxonomic relatedness) that gave rise to such fragments, even if the assignment of these fragments into bins representing the individual species is not possible. Additionally, if specific characteristics of the community are known in advance, such as the likely presence of increased compositional constraints related to restricted environments (i.e., vacuoles within a host cell), the resulting fragment assignments might receive reduced confidence in comparison to the assignments for fragments that arose from environments that lack such constraints.

Conclusions

Collectively, the results presented in this thesis characterize the influence of several factors that influence the distinguishability of microbial genomes. While specific factors, such as fragment heterogeneity or the tendency for a given functional role category to be misclassified may be used to augment existing classifiers as described above, other factors, for instance compositional convergence due to similarities in lifestyle, habitat, or extreme G+C highlight fundamental limitations to the classification of DNA fragments based on compositional characteristics. Despite the fact that the majority of genome pairs considered in this thesis could be distinguished with near-perfect accuracy, many closely related genomes and pairs of genomes that have converged in terms of composition remain nearly (if not completely) indistinguishable. For these difficult cases where the genomes share very similar patterns of genomic

composition, accurate distinguishability on the basis of such patterns is likely to be impossible for short DNA fragments typical of current metagenomic studies. Although distantly related genomes are difficult to distinguish in some instances due to convergence in genome composition, the majority of difficult-to-distinguish genome pairs comprise congeners. In the context of a metagenomic study, the impact of the inability to distinguish between congeners or conspecifics will ultimately depend on the underlying community structure. Many communities may contain congeners that share similar ecological roles, such that studying the ecology of a metagenome relative to higher-level taxonomic groups will still provide valuable insight even if the ecological roles cannot be assigned to specific strains or species within the community. For other communities, however, ecologically distinct strains of the same species may be present [136; 137], and the inability to distinguish between these strains will greatly limit our understanding of such metagenomic communities.

As DNA sequencing technologies inevitably improve, the length of fragments recovered from metagenomic samples is bound to increase, along with the likelihood that such fragments will contain one or more conserved marker genes. When sequencing technologies achieve sufficient read length, it is anticipated that DNA assembly algorithms will allow such reads to be assembled into contigs much longer than is currently possible, even in instances where multiple organisms with similar compositions exist within a community. Furthermore, the association of conserved marker genes with these longer contigs will facilitate binning at more specific taxonomic levels, despite a high degree of similarity in patterns of genome composition. Likewise, an increase in fragment length will also help to mitigate the confounding influence of LGT-derived sequence that has not yet undergone significant amelioration, if such fragments are first examined to identify regions of atypical composition (i.e., using a sequence-segmentation approach as suggested above). In many instances, contigs containing such compositionally atypical sequences will likely be associated with genomic sequence that either contains conserved marker genes or is much more representative of the patterns of genomic composition inherent to the source genome, thus allowing for better discrimination among members of microbial communities.

References

1 Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al.: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**. *Science (80-)* 1995, **269**:496-512.

2 Blattner FR, Plunkett G3, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of Escherichia coli K-12**. *Science (80-)* 1997, **277**:1453-1462.

3 Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA3, Venter JC: **The minimal gene complement of Mycoplasma genitalium**. *Science (80-)* 1995, **270**:397-403.

4 Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessières P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Danchin A, et al.: **The complete genome sequence of the gram-positive bacterium Bacillus subtilis**. *Nature* 1997, **390**:249-256.

5 Staley JT, Konopka A: **Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats**. *Annu Rev Microbiol* 1985, **39**:321-346.

6 Amann RI, Ludwig W, Schleifer KH: **Phylogenetic identification and in situ detection of individual microbial cells without cultivation**. *Microbiol Rev* 1995, **59**:143-169.

7 Torsvik V, Goksøyr J, Daae FL: **High diversity in DNA of soil bacteria**. *Appl Environ Microbiol* 1990, **56**:782-787.

8 Torsvik V, Øvreås L: **Microbial diversity and function in soil: from genes to ecosystems**. *Curr Opin Microbiol* 2002, **5**:240-245.

9 Lorenz P, Liebeton K, Niehaus F, Eck J: **Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space**. *Curr Opin Biotechnol* 2002, **13**:572-577.

10 Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F: **Functional metagenomic profiling of nine biomes**. *Nature* 2008, **452**:629-632.

11 Allen EE, Banfield JF: **Community genomics in microbial ecology and evolution**. *Nat Rev Microbiol* 2005, **3**:489-498.

12 Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S: **Real-time DNA sequencing from single polymerase molecules**. *Science (80-)* 2009, **323**:133-138.

13 Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science (80-)* 2004, **304**:66-74.

14 Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI, Andrews-Pfannkoch C, Fadrosh D, Miller CS, Sutton G, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples**. *PLoS ONE* 2008, **3**:e1456.

15 Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia J, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families**. *PLoS Biol* 2007, **5**:e16.

16 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC: **The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific**. *PLoS Biol* 2007, **5**:e77.

17 Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The human microbiome project**. *Nature* 2007, **449**:804-810.

18 Blackall LL, Crocetti GR, Saunders AM, Bond PL: **A review and update of the microbiology of enhanced biological phosphorus removal in wastewater treatment plants**. *Antonie Van Leeuwenhoek* 2002, **81**:681-691.

19 Bond PL, Erhart R, Wagner M, Keller J, Blackall LL: **Identification of some of the major groups of bacteria in efficient and nonefficient biological phosphorus**

removal activated sludge systems. *Appl Environ Microbiol* 1999, **65**:4077-4084.

20 García Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P: **Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities**. *Nat Biotechnol* 2006, **24**:1263-1269.

21 Hesselmann RP, Werlen C, Hahn D, van der Meer JR, Zehnder AJ: **Enrichment, phylogenetic analysis and detection of a bacterium that performs enhanced biological phosphate removal in activated sludge**. *Syst Appl Microbiol* 1999, **22**:454-465.

22 Crocetti GR, Hugenholtz P, Bond PL, Schuler A, Keller J, Jenkins D, Blackall LL: **Identification of polyphosphate-accumulating organisms and design of 16S rRNA-directed probes for their detection and quantitation**. *Appl Environ Microbiol* 2000, **66**:1175-1182.

23 Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**:37-43.

24 Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative metagenomics of microbial communities**. *Science (80-)* 2005, **308**:554-557.

25 Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, Kyrpides NC: **Use of simulated data sets to evaluate the fidelity of metagenomic processing methods**. *Nat Methods* 2007, **4**:495-500.

26 Eppley JM, Tyson GW, Getz WM, Banfield JF: **Strainer: software for analysis of population variation in community genomic datasets**. *BMC Bioinformatics* 2007, **8**:398.

27 Brulc JM, Antonopoulos DA, Miller MEB, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P, Coutinho PM, Henrissat B, Nelson KE, White BA: **Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases**. *Proc Natl Acad Sci U S A* 2009, **106**:1948-1953.

28 Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms**. *Proc Natl Acad Sci U S A* 1977, **74**:5088-5090.

29 Lloyd AT, Sharp PM: **Evolution of the recA gene and the molecular phylogeny of bacteria**. *J Mol Evol* 1993, **37**:399-407.

30 Woese CR, Maniloff J, Zablen LB: **Phylogenetic analysis of the mycoplasmas**. *Proc Natl Acad Sci U S A* 1980, **77**:494-498.

31 Sandler SJ, Satin LH, Samra HS, Clark AJ: **recA-like genes from three archaean**

species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast Saccharomyces cerevisiae. *Nucleic Acids Res* 1996, **24**:2125-2132.

32 Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML: **Microbial population structures in the deep marine biosphere**. *Science (80-)* 2007, **318**:97-100.

33 Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J: **Phylogenetic classification of short environmental DNA fragments**. *Nucleic Acids Res* 2008, **36**:2230-2239.

34 Rodriguez AA, Bompada T, Syed M, Shah PK, Maltsev N: **Evolutionary analysis of enzymes using Chisel**. *Bioinformatics* 2007, **23**:2961-2968.

35 Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**:377-386.

36 Chan CK, Hsu AL, Halgamuge SK, Tang S: **Binning sequences using very sparse labels within a metagenome**. *BMC Bioinformatics* 2008, **9**:215.

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

38 McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments**. *Nat Methods* 2007, **4**:63-72.

39 Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature**. *Trends Genet* 1995, **11**:283-290.

40 Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO: **Application of tetranucleotide frequencies for the assignment of genomic fragments**. *Environ Microbiol* 2004, **6**:938-947.

41 Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B: **Genomic signature: characterization and classification of species assessed by chaos game representation of sequences**. *Mol Biol Evol* 1999, **16**:1391-1399.

42 Nakashima H, Ota M, Nishikawa K, Ooi T: **Genes from nine genomes are separated into their organisms in the dinucleotide composition space**. *DNA Res* 1998, **5**:251-259.

43 Karlin S, Mrázek J, Campbell AM: **Compositional biases of bacterial genomes and evolutionary implications**. *J Bacteriol* 1997, **179**:3899-3913.

44 Willenbrock H, Friis C, Juncker AS, Ussery DW: **An environmental signature for 323 microbial genomes based on codon adaptation indices**. *Genome Biol* 2006, **7**:R114.

45 Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T: **Informatics for unveiling hidden genome signatures**. *Genome Res* 2003, **13**:693-702.

46 Paul S, Bag SK, Das S, Harvill ET, Dutta C: **Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes**. *Genome Biol* 2008, **9**:R70.

47 Sharp PM, Li WH: **The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications**. *Nucleic Acids Res* 1987, **15**:1281-1295.

48 Foerstner KU, von Mering C, Hooper SD, Bork P: **Environments shape the nucleotide composition of genomes**. *EMBO Rep* 2005, **6**:1208-1213.

49 Inagaki Y, Roger AJ: **Phylogenetic estimation under codon models can be biased by codon usage heterogeneity**. *Mol Phylogenet Evol* 2006, **40**:428-434.

50 Karlin S: **Global dinucleotide signatures and analysis of genomic heterogeneity**. *Curr Opin Microbiol* 1998, **1**:598-610.

51 Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ: **Evolutionary implications of microbial genome tetranucleotide frequency biases**. *Genome Res* 2003, **13**:145-158.

52 Sandberg R, Winberg G, Bränden CI, Kaske A, Ernberg I, Cöster J: **Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier**. *Genome Res* 2001, **11**:1404-1409.

53 Andersson JO: **Lateral gene transfer in eukaryotes**. *Cell Mol Life Sci* 2005, **62**:1182-1197.

54 Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Torres MCM, Giebel JD, Kumar N, Ishmael N, Wang S, Ingram J, Nene RV, Shepard J, Tomkins J, Richards S, Spiro DJ, Ghedin E, Slatko BE, Tettelin H, Werren JH: **Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes**. *Science (80-)* 2007, **317**:1753-1756.

55 Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange**. *J Mol Evol* 1997, **44**:383-397.

56 Ragan MA, Harlow TJ, Beiko RG: **Do different surrogate methods detect lateral genetic transfer events of different relative ages?**. *Trends Microbiol* 2006, **14**:4-8.

57 Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P: **Detection and characterization of horizontal transfers in prokaryotes using genomic signature**. *Nucleic Acids Res* 2005, **33**:e6.

58 Tsirigos A, Rigoutsos I: **A new computational method for the detection of horizontal gene transfer events**. *Nucleic Acids Res* 2005, **33**:922-933.

59 McHardy AC, Rigoutsos I: **What's in the mix: phylogenetic classification of metagenome sequence samples**. *Curr Opin Microbiol* 2007, **10**:499-503.

60 Chen K, Pachter L: **Bioinformatics for whole-genome shotgun sequencing of microbial communities**. *PLoS Comput Biol* 2005, **1**:106-112.

61 Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X,

Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavromatis K, Ivanova N, Kyrpides NC: **The integrated microbial genomes (IMG) system**. *Nucleic Acids Res* 2006, **34**:D344-8.

62 Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B: **Metagenome fragment classification using N-mer frequency profiles**. *Adv Bioinformatics* 2008, **2008**:205969.

63 Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO: **TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences**. *BMC Bioinformatics* 2004, **5**:163.

64 Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T: **Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples**. *DNA Res* 2005, **12**:281-290.

65 Chatterji S, Yamazaki I, Bai Z, Eisen J: **CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads**. In *Research in Computational Molecular Biology*: *2008;* . Edited by ; 2008:17-28.

66 Mavromatis K, Doyle CK, Lykidis A, Ivanova N, Francino MP, Chain P, Shin M, Malfatti S, Larimer F, Copeland A, Detter JC, Land M, Richardson PM, Yu XJ, Walker DH, McBride JW, Kyrpides NC: **The genome of the obligately intracellular bacterium Ehrlichia canis reveals themes of complex membrane structure and immune evasion strategies**. *J Bacteriol* 2006, **188**:4015-4023.

67 Li Y, Jain A: **Classification of Text Documents**. *The Computer Journal* 1998, **41**:537-546.

68 Kohonen J, Talikota S, Corander J, Auvinen P, Arjas E: **A Naive Bayes classifier for protein function prediction**. *In Silico Biol* 2009, **9**:23-34.

69 Androutsopoulos I, Koutsias J, Chandrinos KV, Spyropoulos CD: **An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages**. In : *2000; ACM*. Edited by ; 2000:160-167.

70 Kohonen T: **The self-organizing map**. *Neurocomputing* 1998, **21**:1 - 6.

71 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-410.

72 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

73 Gerlach W, Jünemann S, Tille F, Goesmann A, Stoye J: **WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads**. *BMC Bioinformatics* 2009, **10**:430.

74 Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A: **The Pfam**

protein families database. *Nucleic Acids Res* 2010, **38**:D211-22.

75 Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: **Pfam: multiple sequence alignments and HMM-profiles of protein domains**. *Nucleic Acids Res* 1998, **26**:320-322.

76 Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2)**. *Cladistics* 1989, **5**:164-166.

77 Schmid R, Huson DH: **Contents User Manual for ReadSim V0.7**. 2006, :.

78 Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann K, Krahn I, Krause L, Krömeke H, Kruse O, Mussgnug JH, Neuweger H, Niehaus K, Pühler A, Runte KJ, Szczepanowski R, Tauch A, Tilker A, Viehöver P, Goesmann A: **The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology**. *J Biotechnol* 2008, **136**:77-90.

79 Schbath S: **An efficient statistic to detect over- and under-represented words in DNA sequences**. *J Comput Biol* 1997, **4**:189-192.

80 Mrázek J: **Phylogenetic signals in DNA composition: limitations and prospects**. *Mol Biol Evol* 2009, **26**:1163-1169.

81 Boetius A, Ravenschlag K, Schubert CJ, Rickert D, Widdel F, Gieseke A, Amann R, Jørgensen BB, Witte U, Pfannkuche O: **A marine microbial consortium apparently mediating anaerobic oxidation of methane**. *Nature* 2000, **407**:623-626.

82 Chan CK, Hsu AL, Tang S, Halgamuge SK: **Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing**. *J Biomed Biotechnol* 2008, **2008**:513701.

83 Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, Szeto E, Kyrpides NC, Mussmann M, Amann R, Bergin C, Ruehland C, Rubin EM, Dubilier N: **Symbiosis insights through metagenomic analysis of a microbial consortium**. *Nature* 2006, **443**:950-955.

84 Daffonchio D, Borin S, Frova G, Manachini PL, Sorlini C: **PCR fingerprinting of whole genomes: the spacers between the 16S and 23S rRNA genes and of intergenic tRNA gene regions reveal a different intraspecific genomic variability of Bacillus cereus and Bacillus licheniformis [corrected]**. *Int J Syst Bacteriol* 1998, **48 Pt 1**:107-116.

85 Jolliffe I: *Principal Component Analysis*. Springer, New York, NY; 1986.

86 Wu D, Daugherty SC, Van Aken SE, Pai GH, Watkins KL, Khouri H, Tallon LJ, Zaborsky JM, Dunbar HE, Tran PL, Moran NA, Eisen JA: **Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters**. *PLoS Biol* 2006, **4**:e188.

87 Cambillau C, Claverie JM: **Structural and genomic correlates of hyperthermostability**. *J Biol Chem* 2000, **275**:32383-32386.

88 Suhre K, Claverie J: **Genomic correlates of hyperthermostability, an update**. *J Biol Chem* 2003, **278**:17198-17202.

89 Bohlin J, Skjerve E, Ussery DW: **Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes**. *BMC Genomics* 2008, **9**:104.

90 Audit B, Thermes C, Vaillant C, d'Aubenton-Carafa Y, Muzy JF, Arneodo A: **Long-range correlations in genomic DNA: a signature of the nucleosomal structure**. *Phys Rev Lett* 2001, **86**:2471-2474.

91 Audit B, Vaillant C, Arneodo A, d'Aubenton-Carafa Y, Thermes C: **Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes**. *J Mol Biol* 2002, **316**:903-918.

92 Voss R: **Evolution of long-range fractal correlations and 1/f noise in DNA base sequences**. *Phys Rev Lett* 1992, **68**:3805-3808.

93 Hill KA, Schisler NJ, Singh SM: **Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species**. *J Mol Evol* 1992, **35**:261-269.

94 Geiger DL: **Stretch coding and block coding: two new strategies to represent questionably aligned DNA sequences**. *J Mol Evol* 2002, **54**:191-199.

95 Phillips MJ, Penny D: **The root of the mammalian tree inferred from whole mitochondrial genomes**. *Mol Phylogenet Evol* 2003, **28**:171-185.

96 Campbell W, Campbell J, Reynolds D, Singer E, Torres-Carrasquillo P: **Support vector machines for speaker and language recognition**. *Computer Speech & Language* 2006, **20**:210 - 229.

97 Qi X, Han Y: **Incorporating multiple SVMs for automatic image annotation**. *Pattern Recognit* 2007, **40**:728 - 741.

98 Smirnov DA, Zweitzig DR, Foulk BW, Miller MC, Doyle GV, Pienta KJ, Meropol NJ, Weiner LM, Cohen SJ, Moreno JG, Connelly MC, Terstappen LWMM, O'Hara SM: **Global gene expression profiling of circulating tumor cells**. *Cancer Res* 2005, **65**:4993-4997.

99 Mitra V, Wang C, Banerjee S: **Text classification: A least square support vector machine approach**. *Applied Soft Computing* 2007, **7**:908 - 914.

100 Chang C, Lin C: **LIBSVM: a library for support vector machines**. 2001, :.

101 Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits**. *Bioinformatics* 2008, **24**:319-324.

102 Jordan IK, Rogozin IB, Wolf YI, Koonin EV: **Essential genes are more evolutionarily conserved than are nonessential genes in bacteria**. *Genome Res* 2002, **12**:962-968.

103 Hirsh AE, Fraser HB: **Protein dispensability and rate of evolution**. *Nature*

2001, **411**:1046-1049.

104 Brown KR, Jurisica I: **Online predicted human interaction database**. *Bioinformatics* 2005, **21**:2076-2082.

105 Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM: **The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data**. *Nucleic Acids Res* 2007, **35**:D169-72.

106 R Development Core Team: **R: A Language and Environment for Statistical Computing**. *R Foundation for Statistical Computing, Vienna, Austria* 2008, **ISBN 3-900051-07-0**:URL http://www.R-project.org.

107 Frutos R, Viari A, Ferraz C, Morgat A, Eychenié S, Kandassamy Y, Chantal I, Bensaid A, Coissac E, Vachiery N, Demaille J, Martinez D: **Comparative genomic analysis of three strains of Ehrlichia ruminantium reveals an active process of genome size plasticity**. *J Bacteriol* 2006, **188**:2533-2542.

108 Hotopp JCD, Lin M, Madupu R, Crabtree J, Angiuoli SV, Eisen JA, Seshadri R, Ren Q, Wu M, Utterback TR, Smith S, Lewis M, Khouri H, Zhang C, Niu H, Lin Q, Ohashi N, Zhi N, Nelson W, Brinkac LM, Dodson RJ, Rosovitz MJ, Sundaram J, Daugherty SC, Davidsen T, Durkin AS, Gwinn M, Haft DH, Selengut JD, Sullivan SA, Zafar N, Zhou L, Benahmed F, Forberger H, Halpin R, Mulligan S, Robinson J, White O, Rikihisa Y, Tettelin H: **Comparative genomics of emerging human ehrlichiosis agents**. *PLoS Genet* 2006, **2**:e21.

109 Collins NE, Liebenberg J, de Villiers EP, Brayton KA, Louw E, Pretorius A, Faber FE, van Heerden H, Josemans A, van Kleef M, Steyn HC, van Strijp MF, Zweygarth E, Jongejan F, Maillard JC, Berthier D, Botha M, Joubert F, Corton CH, Thomson NR, Allsopp MT, Allsopp BA: **The genome of the heartwater agent Ehrlichia ruminantium contains multiple tandem repeats of actively variable copy number**. *Proc Natl Acad Sci U S A* 2005, **102**:838-843.

110 McLeod MP, Qin X, Karpathy SE, Gioia J, Highlander SK, Fox GE, McNeill TZ, Jiang H, Muzny D, Jacob LS, Hawes AC, Sodergren E, Gill R, Hume J, Morgan M, Fan G, Amin AG, Gibbs RA, Hong C, Yu X, Walker DH, Weinstock GM: **Complete genome sequence of Rickettsia typhi and comparison with sequences of other rickettsiae**. *J Bacteriol* 2004, **186**:5842-5855.

111 Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie JM, Raoult D: **Mechanisms of evolution in Rickettsia conorii and R. prowazekii**. *Science (80-)* 2001, **293**:2093-2098.

112 Ogata H, Renesto P, Audic S, Robert C, Blanc G, Fournier P, Parinello H, Claverie J, Raoult D: **The genome sequence of Rickettsia felis identifies the first putative conjugative plasmid in an obligate intracellular parasite**. *PLoS Biol* 2005, **3**:e248.

113 Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark

UC, Podowski RM, Näslund AK, Eriksson AS, Winkler HH, Kurland CG: **The genome sequence of Rickettsia prowazekii and the origin of mitochondria**. *Nature* 1998, **396**:133-140.

114 Moran NA: **Microbial minimalism: genome reduction in bacterial pathogens**. *Cell* 2002, **108**:583-586.

115 Renesto P, Ogata H, Audic S, Claverie J, Raoult D: **Some lessons from Rickettsia genomics**. *FEMS Microbiol Rev* 2005, **29**:99-117.

116 Itoh T, Martin W, Nei M: **Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts**. *Proc Natl Acad Sci U S A* 2002, **99**:12944-12948.

117 Blanc G, Ogata H, Robert C, Audic S, Claverie J, Raoult D: **Lateral gene transfer between obligate intracellular bacteria: evidence from the Rickettsia massiliae genome**. *Genome Res* 2007, **17**:1657-1664.

118 Lind PA, Andersson DI: **Whole-genome mutational biases in bacteria**. *Proc Natl Acad Sci U S A* 2008, **105**:17878-17883.

119 Belas R, Horikawa E, Aizawa S, Suvanasuthi R: **Genetic determinants of Silicibacter sp. TM1040 motility**. *J Bacteriol* 2009, **191**:4502-4512.

120 Moran MA, Buchan A, González JM, Heidelberg JF, Whitman WB, Kiene RP, Henriksen JR, King GM, Belas R, Fuqua C, Brinkac L, Lewis M, Johri S, Weaver B, Pai G, Eisen JA, Rahe E, Sheldon WM, Ye W, Miller TR, Carlton J, Rasko DA, Paulsen IT, Ren Q, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Rosovitz MJ, Haft DH, Selengut J, Ward N: **Genome sequence of Silicibacter pomeroyi reveals adaptations to the marine environment**. *Nature* 2004, **432**:910-913.

121 Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW: **Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation**. *Nature* 2003, **424**:1042-1047.

122 Williams KP, Sobral BW, Dickerman AW: **A robust species tree for the alphaproteobacteria**. *J Bacteriol* 2007, **189**:4578-4586.

123 Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, Nishi T, Mori H, Ikemura T: **Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome**. *Gene* 2001, **276**:89-99.

124 Nakamura Y, Itoh T, Matsuda H, Gojobori T: **Biased biological functions of horizontally transferred genes in prokaryotic genomes**. *Nat Genet* 2004, **36**:760-766.

125 van Passel MWJ, Bart A, Thygesen HH, Luyf ACM, van Kampen AHC, van der Ende A: **An acquisition account of genomic islands based on genome signature**

comparisons. *BMC Genomics* 2005, **6**:163.

126 Wang HC, Badger J, Kearney P, Li M: **Analysis of codon usage patterns of bacterial genomes using the self-organizing map**. *Mol Biol Evol* 2001, **18**:792-800.

127 Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF: **Community-wide analysis of microbial genome sequence signatures**. *Genome Biol* 2009, **10**:R85.

128 Karlin S, Ladunga I, Blaisdell BE: **Heterogeneity of genomes: measures and values**. *Proc Natl Acad Sci U S A* 1994, **91**:12837-12841.

129 Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: An information aesthetic for comparative genomics**. *Genome Res* 2009, :.

130 van Passel MWJ, Bart A, Luyf ACM, van Kampen AHC, van der Ende A: **Compositional discordance between prokaryotic plasmids and host chromosomes**. *BMC Genomics* 2006, **7**:26.

131 Coenye T, Vandamme P: **Use of the genomic signature in bacterial classification and identification**. *Syst Appl Microbiol* 2004, **27**:175-185.

132 Bohlin J, Skjerve E, Ussery DW: **Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering**. *BMC Genomics* 2009, **10**:487.

133 Rivera MC, Jain R, Moore JE, Lake JA: **Genomic evidence for two functionally distinct gene classes**. *Proc Natl Acad Sci U S A* 1998, **95**:6239-6244.

134 Keith JM: **Sequence segmentation**. *Methods Mol Biol* 2008, **452**:207-229.

135 Hoff KJ, Lingner T, Meinicke P, Tech M: **Orphelia: predicting genes in metagenomic sequencing reads**. *Nucleic Acids Res* 2009, **37**:W101-5.

136 Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, Post AF, Hagemann M, Paulsen I, Partensky F: **Ecological genomics of marine picocyanobacteria**. *Microbiol Mol Biol Rev* 2009, **73**:249-299.

137 Moran MA, Belas R, Schell MA, González JM, Sun F, Sun S, Binder BJ, Edmonds J, Ye W, Orcutt B, Howard EC, Meile C, Palefsky W, Goesmann A, Ren Q, Paulsen I, Ulrich LE, Thompson LS, Saunders E, Buchan A: **Ecological genomics of marine Roseobacters**. *Appl Environ Microbiol* 2007, **73**:4559-4569.

Appendix 1: List of Genomes Utilized in the Experiments Described in Chapters 3 and 4

All genomes in the following table were utilized in Chapter 4. Genomes labelled with an asterisk were used in Chapter 3.

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Aeropyrum pernix K1	211	Archaea	Crenarchaeota	1.7	56.3
Caldivirga maquilingensis IC-167	17421	Archaea	Crenarchaeota	2.1	43.1
Hyperthermus butylicus DSM 5456	208	Archaea	Crenarchaeota	1.7	53.7
Ignicoccus hospitalis KIN4/I	13914	Archaea	Crenarchaeota	1.3	56.5
Metallosphaera sedula DSM 5348	17447	Archaea	Crenarchaeota	2.2	46.2
Nitrosopumilus maritimus SCM1	19265	Archaea	Crenarchaeota	1.6	34.2
Pyrobaculum aerophilum str. IM2	172	Archaea	Crenarchaeota	2.2	51.4
Pyrobaculum arsenaticum DSM 13514	15582	Archaea	Crenarchaeota	2.1	55.1
Pyrobaculum calidifontis JCM 11548	18111	Archaea	Crenarchaeota	2	57.2
Pyrobaculum islandicum DSM 4184	16743	Archaea	Crenarchaeota	1.8	49.6
Staphylothermus marinus F1	17449	Archaea	Crenarchaeota	1.6	35.7
Sulfolobus acidocaldarius DSM 639	13935	Archaea	Crenarchaeota	2.23	36.7
Sulfolobus solfataricus P2	108	Archaea	Crenarchaeota	3	35.8
Sulfolobus tokodaii str. 7	246	Archaea	Crenarchaeota	2.7	32.8
Thermofilum pendens Hrk 5	16331	Archaea	Crenarchaeota	1.83	57.6
Thermoproteus neutrophilus V24Sta	15645	Archaea	Crenarchaeota	1.8	59.9
Archaeoglobus fulgidus DSM 4304	104	Archaea	Euryarchaeota	2.18	48.6
Candidatus Methanoregula boonei 6A8	18505	Archaea	Euryarchaeota	2.5	54.5
Haloarcula marismortui ATCC 43049	105	Archaea	Euryarchaeota	4.28	61.1
Halobacterium salinarum R1	106	Archaea	Euryarchaeota	2.66	65.7
Halobacterium sp. NRC-1	217	Archaea	Euryarchaeota	2.57	65.9
Haloquadratum walsbyi DSM 16790	17185	Archaea	Euryarchaeota	3.15	47.9
Methanobrevibacter smithii ATCC 35061	18653	Archaea	Euryarchaeota	1.9	31
Methanocaldococcus jannaschii DSM 2661	102	Archaea	Euryarchaeota	1.76	31.3
Methanococcoides burtonii DSM 6242	9634	Archaea	Euryarchaeota	2.58	40.8
Methanococcus aeolicus Nankai-3	18641	Archaea	Euryarchaeota	1.6	30
Methanococcus maripaludis C5	17641	Archaea	Euryarchaeota	1.81	33
Methanococcus maripaludis C6	19639	Archaea	Euryarchaeota	1.7	33.4
Methanococcus maripaludis C7	18819	Archaea	Euryarchaeota	1.8	33.3
Methanococcus maripaludis S2	10632	Archaea	Euryarchaeota	1.66	33.1
Methanococcus vannielii SB	17889	Archaea	Euryarchaeota	1.7	31.3
Methanocorpusculum labreanum Z	18109	Archaea	Euryarchaeota	1.8	50
Methanoculleus marisnigri JR1	16330	Archaea	Euryarchaeota	2.5	62.1

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Methanopyrus kandleri AV19	294	Archaea	Euryarchaeota	1.69	61.2
Methanosaeta thermophila PT	15765	Archaea	Euryarchaeota	1.9	53.5
Methanosarcina acetivorans C2A	290	Archaea	Euryarchaeota	5.75	42.7
Methanosarcina barkeri str. Fusaro	103	Archaea	Euryarchaeota	4.84	39.2
Methanosarcina mazei Go1	300	Archaea	Euryarchaeota	4.1	41.5
Methanosphaera stadtmanae DSM 3091	15579	Archaea	Euryarchaeota	1.77	27.6
Methanospirillum hungatei JF-1	13015	Archaea	Euryarchaeota	3.54	45.1
Methanothermobacter thermautotrophicus str. Delta H	289	Archaea	Euryarchaeota	1.8	49.5
Natronomonas pharaonis DSM 2160	15742	Archaea	Euryarchaeota	2.75	63.1
Picrophilus torridus DSM 9790	10641	Archaea	Euryarchaeota	1.5	36
Pyrococcus abyssi GE5	179	Archaea	Euryarchaeota	1.8	44.7
Pyrococcus furiosus DSM 3638	287	Archaea	Euryarchaeota	1.9	40.8
Pyrococcus horikoshii OT3	207	Archaea	Euryarchaeota	1.7	41.9
Thermococcus kodakarensis KOD1	13213	Archaea	Euryarchaeota	2.09	52
Thermococcus onnurineus NA1	20773	Archaea	Euryarchaeota	1.8	51.3
Thermoplasma acidophilum DSM 1728	110	Archaea	Euryarchaeota	1.6	46
Thermoplasma volcanium GSS1	206	Archaea	Euryarchaeota	1.58	39.9
uncultured methanogenic archaeon RC-I	19641	Archaea	Euryarchaeota	3.2	54.6
Candidatus Korarchaeum cryptofilum OPF8	16525	Archaea	Korarchaeota	1.6	49
Nanoarchaeum equitans Kin4-M	9599	Archaea	Nanoarchaeota	0.49	31.6
Acidobacteria bacterium Ellin345	15771	Bacteria	Acidobacteria	5.7	58.4
Solibacter usitatus Ellin6076	12638	Bacteria	Acidobacteria	10	61.9
Acidothermus cellulolyticus 11B	16097	Bacteria	Actinobacteria	2.4	66.9
Arthrobacter aurescens TC1	12512	Bacteria	Actinobacteria	5.23	62.4
Arthrobacter sp. FB24	12640	Bacteria	Actinobacteria	5.08	65.4
Bifidobacterium adolescentis ATCC 15703	16321	Bacteria	Actinobacteria	2.1	59.2
Bifidobacterium longum DJO10A	18773	Bacteria	Actinobacteria	2.41	60.2
Bifidobacterium longum NCC2705	328	Bacteria	Actinobacteria	2.26	60.1
Bifidobacterium longum subsp. infantis ATCC 15697	17189	Bacteria	Actinobacteria	2.8	59.9
Clavibacter michiganensis subsp. michiganensis NCPPB 382	19643	Bacteria	Actinobacteria	3.4	72.5
Clavibacter michiganensis subsp. sepedonicus	184	Bacteria	Actinobacteria	3.44	72.4
Corynebacterium diphtheriae NCTC 13129	87	Bacteria	Actinobacteria	2.49	53.5
Corynebacterium efficiens YS-314	305	Bacteria	Actinobacteria	3.1	63.1
Corynebacterium glutamicum ATCC 13032	307	Bacteria	Actinobacteria	3.3	53.8
Corynebacterium glutamicum ATCC 13032 DSM 20300	13760	Bacteria	Actinobacteria	3.3	53.8
Corynebacterium glutamicum R	19193	Bacteria	Actinobacteria	3.35	54.1
Corynebacterium jeikeium K411	13967	Bacteria	Actinobacteria	2.51	61.4

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Corynebacterium urealyticum DSM 7109	29211	Bacteria	Actinobacteria	2.4	64.2
Frankia alni ACN14a	17403	Bacteria	Actinobacteria	7.5	72.8
Frankia sp. CcI3	13963	Bacteria	Actinobacteria	5.4	70.1
Frankia sp. EAN1pec	13915	Bacteria	Actinobacteria	9	71.2
Kineococcus radiotolerans SRS30216	10689	Bacteria	Actinobacteria	4.99	74.2
Kocuria rhizophila DC2201	27833	Bacteria	Actinobacteria	2.7	71.2
Leifsonia xyli subsp. xyli str. CTCB07	212	Bacteria	Actinobacteria	2.58	67.7
Mycobacterium abscessus	15691	Bacteria	Actinobacteria	5.12	64.1
Mycobacterium avium 104	88	Bacteria	Actinobacteria	5.5	69
Mycobacterium avium subsp. paratuberculosis K-10	91	Bacteria	Actinobacteria	4.8	69.3
Mycobacterium bovis AF2122/97	89	Bacteria	Actinobacteria	4.35	65.6
Mycobacterium bovis BCG str. Pasteur 1173P2	18059	Bacteria	Actinobacteria	4.4	65.6
Mycobacterium gilvum PYR-GCK	15760	Bacteria	Actinobacteria	5.96	67.7
Mycobacterium leprae TN	90	Bacteria	Actinobacteria	3.27	57.8
Mycobacterium marinum M	16725	Bacteria	Actinobacteria	6.62	65.7
Mycobacterium smegmatis str. MC2 155	92	Bacteria	Actinobacteria	7	67.4
Mycobacterium sp. JLS	16079	Bacteria	Actinobacteria	6	68.4
Mycobacterium sp. KMS	16081	Bacteria	Actinobacteria	6.22	68.2
Mycobacterium sp. MCS	15762	Bacteria	Actinobacteria	5.92	68.4
Mycobacterium tuberculosis CDC1551	223	Bacteria	Actinobacteria	4.4	65.6
Mycobacterium tuberculosis F11	15642	Bacteria	Actinobacteria	4.4	65.6
Mycobacterium tuberculosis H37Ra	18883	Bacteria	Actinobacteria	4.4	65.6
Mycobacterium tuberculosis H37Rv	224	Bacteria	Actinobacteria	4.4	65.6
Mycobacterium ulcerans Agy99	16230	Bacteria	Actinobacteria	5.77	65.4
Mycobacterium vanbaalenii PYR-1	15761	Bacteria	Actinobacteria	6.5	67.8
Nocardia farcinica IFM 10152	13117	Bacteria	Actinobacteria	6.29	70.7
Nocardioides sp. JS614	12738	Bacteria	Actinobacteria	5.31	71.4
Propionibacterium acnes KPA171202	12460	Bacteria	Actinobacteria	2.56	60
Renibacterium salmoninarum ATCC 33209	19227	Bacteria	Actinobacteria	3.2	56.3
Rhodococcus jostii RHA1	13693	Bacteria	Actinobacteria	9.67	67
Rubrobacter xylanophilus DSM 9941	10670	Bacteria	Actinobacteria	3.23	70.5
Saccharopolyspora erythraea NRRL 2338	18489	Bacteria	Actinobacteria	8.2	71.1
Salinispora arenicola CNS-205	17109	Bacteria	Actinobacteria	5.8	69.5
Salinispora tropica CNB-440	16342	Bacteria	Actinobacteria	5.2	69.5
Streptomyces avermitilis MA-4680	189	Bacteria	Actinobacteria	9.09	70.7
Streptomyces coelicolor A3(2)	242	Bacteria	Actinobacteria	9.09	72
Streptomyces griseus subsp. griseus NBRC 13350	20085	Bacteria	Actinobacteria	8.5	72.2
Thermobifida fusca YX	94	Bacteria	Actinobacteria	3.6	67.5

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Tropheryma whipplei str. Twist	95	Bacteria	Actinobacteria	0.93	46.3
Tropheryma whipplei TW08/27	354	Bacteria	Actinobacteria	0.93	46.3
Aquifex aeolicus VF5	215	Bacteria	Aquificae	1.59	43.3
Hydrogenobaculum sp. Y04AAS1	18891	Bacteria	Aquificae	1.6	34.8
Sulfurihydrogenibium sp. YO3AOP1	18889	Bacteria	Aquificae	1.8	32
Bacteroides fragilis NCTC 9343	46	Bacteria	Bacteroidetes	5.24	43.1
Bacteroides fragilis YCH46	13067	Bacteria	Bacteroidetes	5.31	43.2
Bacteroides thetaiotaomicron VPI-5482	399	Bacteria	Bacteroidetes	6.33	42.9
Bacteroides vulgatus ATCC 8482	13378	Bacteria	Bacteroidetes	5.2	42.2
Candidatus Amoebophilus asiaticus 5a2	19981	Bacteria	Bacteroidetes	1.9	35
Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	29025	Bacteria	Bacteroidetes	1.21	32.9
Candidatus Sulcia muelleri GWSS	19617	Bacteria	Bacteroidetes	0.25	22.4
Cytophaga hutchinsonii ATCC 33406	54	Bacteria	Bacteroidetes	4.4	38.8
Flavobacterium johnsoniae UW101	16082	Bacteria	Bacteroidetes	6.1	34.1
Flavobacterium psychrophilum JIP02/86	19979	Bacteria	Bacteroidetes	2.9	32.5
Gramella forsetii KT0803	19061	Bacteria	Bacteroidetes	3.8	36.6
Parabacteroides distasonis ATCC 8503	13485	Bacteria	Bacteroidetes	4.8	45.1
Porphyromonas gingivalis ATCC 33277	19051	Bacteria	Bacteroidetes	2.4	48.4
Porphyromonas gingivalis W83	48	Bacteria	Bacteroidetes	2.34	48.3
Salinibacter ruber DSM 13855	16159	Bacteria	Bacteroidetes	3.59	66.1
Elusimicrobium minutum Pei191	19701	Bacteria	candidate division TG1	1.6	40
Candidatus Protochlamydia amoebophila UWE25	10700	Bacteria	Chlamydiae	2.41	34.7
Chlamydia muridarum Nigg	229	Bacteria	Chlamydiae	1.08	40.3
Chlamydia trachomatis 434/Bu	28583	Bacteria	Chlamydiae	1	41.3
Chlamydia trachomatis A/HAR-13	13885	Bacteria	Chlamydiae	1.01	41.3
Chlamydia trachomatis D/UW-3/CX	45	Bacteria	Chlamydiae	1.04	41.3
Chlamydia trachomatis L2b/UCH-1/proctitis	28585	Bacteria	Chlamydiae	1	41.3
Chlamydophila abortus S26/3	355	Bacteria	Chlamydiae	1.14	39.9
Chlamydophila caviae GPIC	228	Bacteria	Chlamydiae	1.18	39.2
Chlamydophila felis Fe/C-56	370	Bacteria	Chlamydiae	1.21	39.3
Chlamydophila pneumoniae AR39	247	Bacteria	Chlamydiae	1.23	40.6
Chlamydophila pneumoniae CWL029	248	Bacteria	Chlamydiae	1.2	40.6
Chlamydophila pneumoniae J138	257	Bacteria	Chlamydiae	1.2	40.6
Chlamydophila pneumoniae TW-183	420	Bacteria	Chlamydiae	1.23	40.6
Chlorobaculum parvum NCIB 8327	29213	Bacteria	Chlorobi	2.3	55.8
Chlorobium chlorochromatii CaD3	13921	Bacteria	Chlorobi	2.6	44.3
Chlorobium limicola DSM 245	12606	Bacteria	Chlorobi	2.8	51.3
Chlorobium phaeobacteroides BS1	12608	Bacteria	Chlorobi	2.7	48.9

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Chlorobium phaeobacteroides DSM 266	12609	Bacteria	Chlorobi	3.1	48.4
Chlorobium phaeovibrioides DSM 265	12607	Bacteria	Chlorobi	2	53
Chlorobium tepidum TLS	302	Bacteria	Chlorobi	2.2	56.5
Chloroherpeton thalassium ATCC 35110	29215	Bacteria	Chlorobi	3.3	45
Pelodictyon luteolum DSM 273	13012	Bacteria	Chlorobi	2.36	57.3
Pelodictyon phaeoclathratiforme BU-1	13011	Bacteria	Chlorobi	3	48.1
Prosthecochloris aestuarii DSM 271	12749	Bacteria	Chlorobi	2.57	50.1
Chloroflexus aurantiacus J-10-fl	59	Bacteria	Chloroflexi	5.3	56.7
Dehalococcoides ethenogenes 195	214	Bacteria	Chloroflexi	1.47	48.9
Dehalococcoides sp. BAV1	15770	Bacteria	Chloroflexi	1.3	47.2
Dehalococcoides sp. CBDB1	15604	Bacteria	Chloroflexi	1.4	47
Herpetosiphon aurantiacus ATCC 23779	16523	Bacteria	Chloroflexi	6.74	50.9
Roseiflexus castenholzii DSM 13941	13462	Bacteria	Chloroflexi	5.7	60.7
Roseiflexus sp. RS-1	16190	Bacteria	Chloroflexi	5.8	60.4
Acaryochloris marina MBIC11017	12997	Bacteria	Cyanobacteria	8.36	47
Anabaena variabilis ATCC 29413	10642	Bacteria	Cyanobacteria	7.07	41.4
Cyanothece sp. ATCC 51142	20319	Bacteria	Cyanobacteria	5.43	37.9
Gloeobacter violaceus PCC 7421	9606	Bacteria	Cyanobacteria	4.66	62
Microcystis aeruginosa NIES-843	27835	Bacteria	Cyanobacteria	5.8	42.3
Nostoc punctiforme PCC 73102	216	Bacteria	Cyanobacteria	9.01	41.4
Nostoc sp. PCC 7120	244	Bacteria	Cyanobacteria	7.21	41.3
Prochlorococcus marinus str. AS9601	13548	Bacteria	Cyanobacteria	1.7	31.3
Prochlorococcus marinus str. MIT 9211	13551	Bacteria	Cyanobacteria	1.7	38
Prochlorococcus marinus str. MIT 9215	18633	Bacteria	Cyanobacteria	1.7	31.1
Prochlorococcus marinus str. MIT 9301	15746	Bacteria	Cyanobacteria	1.6	31.3
Prochlorococcus marinus str. MIT 9303	13496	Bacteria	Cyanobacteria	2.7	50
Prochlorococcus marinus str. MIT 9312	13910	Bacteria	Cyanobacteria	1.71	31.2
Prochlorococcus marinus str. MIT 9313	220	Bacteria	Cyanobacteria	2.41	50.7
Prochlorococcus marinus str. MIT 9515	13617	Bacteria	Cyanobacteria	1.7	30.8
Prochlorococcus marinus str. NATL1A	15660	Bacteria	Cyanobacteria	1.9	35
Prochlorococcus marinus str. NATL2A	13911	Bacteria	Cyanobacteria	1.8	35.1
Prochlorococcus marinus subsp. marinus str. CCMP1375	419	Bacteria	Cyanobacteria	1.75	36.4
Prochlorococcus marinus subsp. pastoris str. CCMP1986	213	Bacteria	Cyanobacteria	1.7	30.8
Synechococcus elongatus PCC 6301	13282	Bacteria	Cyanobacteria	2.7	55.5
Synechococcus elongatus PCC 7942	10645	Bacteria	Cyanobacteria	2.75	55.4
Synechococcus sp. CC9311	12530	Bacteria	Cyanobacteria	2.61	52.4
Synechococcus sp. CC9605	13643	Bacteria	Cyanobacteria	2.51	59.2
Synechococcus sp. CC9902	13655	Bacteria	Cyanobacteria	2.2	54.2

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Synechococcus sp. JA-2-3Ba(2-13)	16252	Bacteria	Cyanobacteria	3	58.5
Synechococcus sp. JA-3-3Ab	16251	Bacteria	Cyanobacteria	2.9	60.2
Synechococcus sp. PCC 7002	28247	Bacteria	Cyanobacteria	3.4	49.2
Synechococcus sp. RCC307	13654	Bacteria	Cyanobacteria	2.2	60.8
Synechococcus sp. WH 7803	13642	Bacteria	Cyanobacteria	2.4	60.2
Synechococcus sp. WH 8102	230	Bacteria	Cyanobacteria	2.43	59.4
Synechocystis sp. PCC 6803	60	Bacteria	Cyanobacteria	3.94	47.4
Thermosynechococcus elongatus BP-1	308	Bacteria	Cyanobacteria	2.59	53.9
Trichodesmium erythraeum IMS101	318	Bacteria	Cyanobacteria	7.8	34.1
Deinococcus geothermalis DSM 11300	13423	Bacteria	Deinococcus-Thermus	3.28	66.5
Deinococcus radiodurans R1	65	Bacteria	Deinococcus-Thermus	3.24	66.6
Thermus thermophilus HB27	10617	Bacteria	Deinococcus-Thermus	2.13	69.4
Thermus thermophilus HB8	13202	Bacteria	Deinococcus-Thermus	2.07	69.5
Dictyoglomus thermophilum H-6-12	30731	Bacteria	Dictyoglomi	2	33.7
Alkaliphilus metalliredigens QYMF	13006	Bacteria	Firmicutes	4.9	36.8
Alkaliphilus oremlandii OhILAs	16083	Bacteria	Firmicutes	3.1	36.3
Anoxybacillus flavithermus WK1	28245	Bacteria	Firmicutes	2.8	41.8
Bacillus amyloliquefaciens FZB42	13403	Bacteria	Firmicutes	3.9	46.5
Bacillus anthracis str. Ames	309	Bacteria	Firmicutes	5.23	35.4
Bacillus anthracis str. Ames Ancestor	10784	Bacteria	Firmicutes	5.47	35.2
Bacillus anthracis str. Sterne	10878	Bacteria	Firmicutes	5.23	35.4
Bacillus cereus ATCC 10987	74	Bacteria	Firmicutes	5.43	35.5
Bacillus cereus ATCC 14579	384	Bacteria	Firmicutes	5.42	35.3
Bacillus cereus E33L	12468	Bacteria	Firmicutes	5.85	35.1
Bacillus cereus subsp. cytotoxis NVH 391-98	13624	Bacteria	Firmicutes	4.11	35.9
Bacillus clausii KSM-K16	13291	Bacteria	Firmicutes	4.3	44.8
Bacillus halodurans C-125	235	Bacteria	Firmicutes	4.2	43.7
Bacillus licheniformis ATCC 14580; DSM 13	12388	Bacteria	Firmicutes	4.2	46.2
Bacillus licheniformis DSM 13; ATCC 14580	13082	Bacteria	Firmicutes	4.2	46.2
Bacillus pumilus SAFR-032	20391	Bacteria	Firmicutes	3.7	41.3
Bacillus subtilis subsp. subtilis str. 168	76	Bacteria	Firmicutes	4.2	43.5
Bacillus thuringiensis serovar konkukian str. 97-27	10877	Bacteria	Firmicutes	5.28	35.4
Bacillus thuringiensis str. Al Hakam	18255	Bacteria	Firmicutes	5.36	35.4
Bacillus weihenstephanensis KBAB4	13623	Bacteria	Firmicutes	5.91	35.5
Caldicellulosiruptor saccharolyticus DSM 8903	13466	Bacteria	Firmicutes	3	35.3

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Candidatus Desulforudis audaxviator MP104C	21047	Bacteria	Firmicutes	2.3	60.8
Carboxydothermus hydrogenoformans Z-2901	253	Bacteria	Firmicutes	2.4	42
Clostridium acetobutylicum ATCC 824	77	Bacteria	Firmicutes	4.13	30.9
Clostridium beijerinckii NCIMB 8052	12637	Bacteria	Firmicutes	6	29.9
Clostridium botulinum A str. ATCC 19397	19517	Bacteria	Firmicutes	3.9	28.2
Clostridium botulinum A str. ATCC 3502	193	Bacteria	Firmicutes	3.92	28.2
Clostridium botulinum A str. Hall	19521	Bacteria	Firmicutes	3.8	28.2
Clostridium botulinum A3 str. Loch Maree	28507	Bacteria	Firmicutes	4.27	28.1
Clostridium botulinum B str. Eklund 17B	28857	Bacteria	Firmicutes	3.85	27.5
Clostridium botulinum B1 str. Okra	28505	Bacteria	Firmicutes	4.15	28.2
Clostridium botulinum E3 str. Alaska E43	28855	Bacteria	Firmicutes	3.7	27.4
Clostridium botulinum F str. Langeland	19519	Bacteria	Firmicutes	4.02	28.3
Clostridium difficile 630	78	Bacteria	Firmicutes	4.31	29.1
Clostridium kluyveri DSM 555	19065	Bacteria	Firmicutes	4.06	32
Clostridium novyi NT	16820	Bacteria	Firmicutes	2.5	28.9
Clostridium perfringens ATCC 13124	304	Bacteria	Firmicutes	3.3	28.4
Clostridium perfringens SM101	12521	Bacteria	Firmicutes	2.92	28.2
Clostridium perfringens str. 13	79	Bacteria	Firmicutes	3.05	28.5
Clostridium phytofermentans ISDg	16184	Bacteria	Firmicutes	4.8	35.3
Clostridium tetani E88	81	Bacteria	Firmicutes	2.87	28.6
Clostridium thermocellum ATCC 27405	314	Bacteria	Firmicutes	3.8	39
Coprothermobacter proteolyticus DSM 5265	30729	Bacteria	Firmicutes	1.4	44.8
Desulfitobacterium hafniense Y51	16639	Bacteria	Firmicutes	5.73	47.4
Desulfotomaculum reducens MI-1	13424	Bacteria	Firmicutes	3.6	42.3
Enterococcus faecalis V583	70	Bacteria	Firmicutes	3.36	37.4
Exiguobacterium sibiricum 255-15	10649	Bacteria	Firmicutes	3.01	47.7
Finegoldia magna ATCC 29328	18981	Bacteria	Firmicutes	1.99	32.1
Geobacillus kaustophilus HTA426	13233	Bacteria	Firmicutes	3.59	52
Geobacillus thermodenitrificans NG80-2	18655	Bacteria	Firmicutes	3.66	48.9
Heliobacterium modesticaldum Ice1	13427	Bacteria	Firmicutes	3.1	57
Lactobacillus acidophilus NCFM	82	Bacteria	Firmicutes	2	34.7
Lactobacillus brevis ATCC 367	404	Bacteria	Firmicutes	2.35	46.1
Lactobacillus casei ATCC 334	402	Bacteria	Firmicutes	2.93	46.6
Lactobacillus casei BL23	30359	Bacteria	Firmicutes	3.1	46.3
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	16871	Bacteria	Firmicutes	1.9	49.7
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	403	Bacteria	Firmicutes	1.9	49.7
Lactobacillus fermentum IFO 3956	18979	Bacteria	Firmicutes	2.1	51.5

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Lactobacillus gasseri ATCC 33323	84	Bacteria	Firmicutes	1.9	35.3
Lactobacillus helveticus DPC 4571	17811	Bacteria	Firmicutes	2.1	37.1
Lactobacillus johnsonii NCC 533	9638	Bacteria	Firmicutes	2	34.6
Lactobacillus plantarum WCFS1	356	Bacteria	Firmicutes	3.34	44.4
Lactobacillus reuteri DSM 20016	15766	Bacteria	Firmicutes	2	38.9
Lactobacillus reuteri JCM 1112	19011	Bacteria	Firmicutes	2	38.9
Lactobacillus sakei subsp. sakei 23K	13435	Bacteria	Firmicutes	1.9	41.3
Lactobacillus salivarius UCC118	13280	Bacteria	Firmicutes	2.1	33
Lactococcus lactis subsp. cremoris MG1363	18797	Bacteria	Firmicutes	2.5	35.7
Lactococcus lactis subsp. cremoris SK11	401	Bacteria	Firmicutes	2.56	35.8
Lactococcus lactis subsp. lactis Il1403	72	Bacteria	Firmicutes	2.4	35.3
Leuconostoc citreum KM20	16062	Bacteria	Firmicutes	1.9	38.9
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	315	Bacteria	Firmicutes	2.04	37.7
Listeria innocua Clip11262	86	Bacteria	Firmicutes	3.09	37.4
Listeria monocytogenes EGD-e	276	Bacteria	Firmicutes	2.94	38
Listeria monocytogenes str. 4b F2365	85	Bacteria	Firmicutes	2.91	38
Listeria welshimeri serovar 6b str. SLCC5334	13443	Bacteria	Firmicutes	2.8	36.4
Lysinibacillus sphaericus C3-41	19619	Bacteria	Firmicutes	4.78	37.1
Moorella thermoacetica ATCC 39073	10648	Bacteria	Firmicutes	2.6	55.8
Natranaerobius thermophilus JW/NM-WN-LF	20207	Bacteria	Firmicutes	3.23	36.3
Oceanobacillus iheyensis HTE831	284	Bacteria	Firmicutes	3.63	35.7
Oenococcus oeni PSU-1	317	Bacteria	Firmicutes	1.8	37.9
Pediococcus pentosaceus ATCC 25745	398	Bacteria	Firmicutes	1.8	37.4
Pelotomaculum thermopropionicum SI	19023	Bacteria	Firmicutes	3	53
Staphylococcus aureus RF122	63	Bacteria	Firmicutes	2.7	32.8
Staphylococcus aureus subsp. aureus COL	238	Bacteria	Firmicutes	2.8	32.8
Staphylococcus aureus subsp. aureus JH1	15758	Bacteria	Firmicutes	2.93	32.9
Staphylococcus aureus subsp. aureus JH9	15757	Bacteria	Firmicutes	2.93	32.9
Staphylococcus aureus subsp. aureus MRSA252	265	Bacteria	Firmicutes	2.9	32.8
Staphylococcus aureus subsp. aureus MSSA476	266	Bacteria	Firmicutes	2.82	32.8
Staphylococcus aureus subsp. aureus Mu3	18509	Bacteria	Firmicutes	2.9	32.9
Staphylococcus aureus subsp. aureus Mu50	263	Bacteria	Firmicutes	2.93	32.8
Staphylococcus aureus subsp. aureus MW2	306	Bacteria	Firmicutes	2.8	32.8
Staphylococcus aureus subsp. aureus N315	264	Bacteria	Firmicutes	2.82	32.8
Staphylococcus aureus subsp. aureus NCTC 8325	237	Bacteria	Firmicutes	2.8	32.9
Staphylococcus aureus subsp. aureus str. Newman	18801	Bacteria	Firmicutes	2.9	32.9

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Staphylococcus aureus subsp. aureus USA300	16313	Bacteria	Firmicutes	2.94	32.7
Staphylococcus aureus subsp. aureus USA300_TCH1516	19489	Bacteria	Firmicutes	2.93	32.7
Staphylococcus epidermidis ATCC 12228	279	Bacteria	Firmicutes	2.56	32
Staphylococcus epidermidis RP62A	64	Bacteria	Firmicutes	2.64	32.1
Staphylococcus haemolyticus JCSC1435	12508	Bacteria	Firmicutes	* 2.7	32.8
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	15596	Bacteria	Firmicutes	2.56	33.2
Streptococcus agalactiae 2603V/R	330	Bacteria	Firmicutes	2.2	35.6
Streptococcus agalactiae A909	326	Bacteria	Firmicutes	2.13	35.6
Streptococcus agalactiae NEM316	334	Bacteria	Firmicutes	2.2	35.6
Streptococcus equi subsp. zooepidemicus MGCS10565	30781	Bacteria	Firmicutes	2	41.8
Streptococcus gordonii str. Challis substr. CH1	66	Bacteria	Firmicutes	2.2	40.5
Streptococcus mutans UA159	333	Bacteria	Firmicutes	2.03	36.8
Streptococcus pneumoniae CGSP14	29179	Bacteria	Firmicutes	2.2	39.5
Streptococcus pneumoniae D39	16374	Bacteria	Firmicutes	2	39.7
Streptococcus pneumoniae G54	29047	Bacteria	Firmicutes	2.1	39.6
Streptococcus pneumoniae Hungary19A-6	28035	Bacteria	Firmicutes	2.2	39.6
Streptococcus pneumoniae R6	278	Bacteria	Firmicutes	2.04	39.7
Streptococcus pneumoniae TIGR4	277	Bacteria	Firmicutes	2.2	39.7
Streptococcus pyogenes M1 GAS	269	Bacteria	Firmicutes	1.9	38.5
Streptococcus pyogenes MGAS10270	16364	Bacteria	Firmicutes	1.9	38.4
Streptococcus pyogenes MGAS10394	12469	Bacteria	Firmicutes	1.9	38.7
Streptococcus pyogenes MGAS10750	16366	Bacteria	Firmicutes	1.9	38.3
Streptococcus pyogenes MGAS2096	16365	Bacteria	Firmicutes	1.9	38.7
Streptococcus pyogenes MGAS315	311	Bacteria	Firmicutes	1.9	38.6
Streptococcus pyogenes MGAS5005	13888	Bacteria	Firmicutes	1.8	38.5
Streptococcus pyogenes MGAS6180	13887	Bacteria	Firmicutes	1.9	38.4
Streptococcus pyogenes MGAS8232	286	Bacteria	Firmicutes	1.9	38.5
Streptococcus pyogenes MGAS9429	16363	Bacteria	Firmicutes	1.8	38.5
Streptococcus pyogenes NZ131	20707	Bacteria	Firmicutes	1.8	38.6
Streptococcus pyogenes SSI-1	301	Bacteria	Firmicutes	1.9	38.6
Streptococcus pyogenes str. Manfredo	270	Bacteria	Firmicutes	1.8	38.6
Streptococcus sanguinis SK36	13942	Bacteria	Firmicutes	2.4	43.4
Streptococcus suis 05ZYH33	17153	Bacteria	Firmicutes	2.1	41.1
Streptococcus suis 98HAH33	17155	Bacteria	Firmicutes	2.1	41.1
Streptococcus thermophilus CNRZ1066	13163	Bacteria	Firmicutes	1.8	39.1
Streptococcus thermophilus LMD-9	13773	Bacteria	Firmicutes	1.91	39.1
Streptococcus thermophilus LMG 18311	13162	Bacteria	Firmicutes	1.8	39.1

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Symbiobacterium thermophilum IAM 14863	12994	Bacteria	Firmicutes	3.6	68.7
Syntrophomonas wolfei subsp. wolfei str. Goettingen	13014	Bacteria	Firmicutes	2.94	44.9
Thermoanaerobacter pseudethanolicus ATCC 33223	13901	Bacteria	Firmicutes	2.4	34.5
Thermoanaerobacter sp. X514	16394	Bacteria	Firmicutes	2.5	34.5
Thermoanaerobacter tengcongensis MB4	249	Bacteria	Firmicutes	2.69	37.6
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	295	Bacteria	Fusobacteria	2.17	27.2
Thermodesulfovibrio yellowstonii DSM 11347	30733	Bacteria	Nitrospirae	2	34.1
Rhodopirellula baltica SH 1	413	Bacteria	Planctomycetes	7.1	55.4
Acidiphilium cryptum JF-5	15753	Bacteria	Proteobacteria	3.97	67.1
Acidithiobacillus ferrooxidans ATCC 53993	16689	Bacteria	Proteobacteria	2.9	58.9
Acidovorax avenae subsp. citrulli AAC00-1	15708	Bacteria	Proteobacteria	5.4	68.5
Acidovorax sp. JS42	15685	Bacteria	Proteobacteria	4.54	66.1
Acinetobacter baumannii AB0057	21111	Bacteria	Proteobacteria	4.11	39.2
Acinetobacter baumannii ACICU	17827	Bacteria	Proteobacteria	3.99	38.9
Acinetobacter baumannii ATCC 17978	17477	Bacteria	Proteobacteria	4.02	38.9
Acinetobacter baumannii AYE	28921	Bacteria	Proteobacteria	4.01	39.3
Acinetobacter baumannii SDF	13001	Bacteria	Proteobacteria	3.46	39.1
Acinetobacter sp. ADP1	12352	Bacteria	Proteobacteria	3.6	40.4
Actinobacillus pleuropneumoniae L20	18221	Bacteria	Proteobacteria	2.3	41.3
Actinobacillus pleuropneumoniae serovar 3 str. JL03	19135	Bacteria	Proteobacteria	2.2	41.2
Actinobacillus pleuropneumoniae serovar 7 str. AP76	29909	Bacteria	Proteobacteria	2.31	41.2
Actinobacillus succinogenes 130Z	13370	Bacteria	Proteobacteria	2.3	44.9
Aeromonas hydrophila subsp. hydrophila ATCC 7966	16697	Bacteria	Proteobacteria	4.7	61.5
Aeromonas salmonicida subsp. salmonicida A449	16723	Bacteria	Proteobacteria	5.05	58.2
*Agrobacterium tumefaciens str. C58	283	Bacteria	Proteobacteria	5.65	59
Alcanivorax borkumensis SK2	13005	Bacteria	Proteobacteria	3.1	54.7
Aliivibrio salmonicida LFI1238	30703	Bacteria	Proteobacteria	4.62	39
Alkalilimnicola ehrlichei MLHE-1	15763	Bacteria	Proteobacteria	3.3	67.5
Alteromonas macleodii Deep ecotype	13374	Bacteria	Proteobacteria	4.4	44.9
Anaeromyxobacter dehalogenans 2CP-C	12634	Bacteria	Proteobacteria	5	74.9
Anaeromyxobacter sp. Fw109-5	17729	Bacteria	Proteobacteria	5.3	73.5
Anaeromyxobacter sp. K	19743	Bacteria	Proteobacteria	5.1	74.8
*Anaplasma marginale str. St. Maries	40	Bacteria	Proteobacteria	1.2	49.8
*Anaplasma phagocytophilum HZ	336	Bacteria	Proteobacteria	1.47	41.6
Arcobacter butzleri RM4018	16319	Bacteria	Proteobacteria	2.3	27
Aromatoleum aromaticum EbN1	13242	Bacteria	Proteobacteria	4.73	64.7

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Azoarcus sp. BH72	13217	Bacteria	Proteobacteria	4.4	67.9
Azorhizobium caulinodans ORS 571	19267	Bacteria	Proteobacteria	5.4	67.3
*Bartonella bacilliformis KC583	16249	Bacteria	Proteobacteria	1.4	38.2
*Bartonella henselae str. Houston-1	196	Bacteria	Proteobacteria	1.93	38.2
*Bartonella quintana str. Toulouse	44	Bacteria	Proteobacteria	1.58	38.8
Bartonella tribocorum CIP 105476	28109	Bacteria	Proteobacteria	2.62	38.8
Baumannia cicadellinicola str. Hc (Homalodisca coagulata)	12513	Bacteria	Proteobacteria	0.69	33.2
Bdellovibrio bacteriovorus HD100	9637	Bacteria	Proteobacteria	3.8	50.6
Beijerinckia indica subsp. indica ATCC 9039	20841	Bacteria	Proteobacteria	4.45	57
Bordetella avium 197N	27	Bacteria	Proteobacteria	3.7	61.6
Bordetella bronchiseptica RB50	24	Bacteria	Proteobacteria	5.3	68.1
Bordetella parapertussis 12822	25	Bacteria	Proteobacteria	4.77	68.1
Bordetella pertussis Tohama I	26	Bacteria	Proteobacteria	4.1	67.7
Bordetella petrii DSM 12804	28135	Bacteria	Proteobacteria	5.3	65.5
*Bradyrhizobium japonicum USDA 110	17	Bacteria	Proteobacteria	9.1	64.1
Bradyrhizobium sp. BTAi1	16137	Bacteria	Proteobacteria	8.53	64.8
Bradyrhizobium sp. ORS278	19575	Bacteria	Proteobacteria	7.5	65.5
*Brucella abortus bv. 1 str. 9-941	9619	Bacteria	Proteobacteria	3.3	57.2
Brucella abortus S19	18999	Bacteria	Proteobacteria	3.3	57.2
Brucella canis ATCC 23365	20243	Bacteria	Proteobacteria	3.3	57.2
*Brucella melitensis 16M	180	Bacteria	Proteobacteria	3.29	57.2
*Brucella melitensis biovar Abortus 2308	16203	Bacteria	Proteobacteria	3.32	57.2
Brucella ovis ATCC 25840	12514	Bacteria	Proteobacteria	3.3	57.2
*Brucella suis 1330	320	Bacteria	Proteobacteria	3.31	57.3
Brucella suis ATCC 23445	20371	Bacteria	Proteobacteria	3.3	57.2
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	245	Bacteria	Proteobacteria	0.66	26.4
Buchnera aphidicola str. Bp (Baizongia pistaciae)	256	Bacteria	Proteobacteria	0.62	25.3
Buchnera aphidicola str. Cc (Cinara cedri)	16372	Bacteria	Proteobacteria	0.42	20.2
Buchnera aphidicola str. Sg (Schizaphis graminum)	312	Bacteria	Proteobacteria	0.64	25.3
Burkholderia ambifaria AMMD	13490	Bacteria	Proteobacteria	7.57	66.8
Burkholderia ambifaria MC40-6	17411	Bacteria	Proteobacteria	7.6	66.4
Burkholderia cenocepacia AU 1054	13919	Bacteria	Proteobacteria	7.28	66.9
Burkholderia cenocepacia HI2424	13918	Bacteria	Proteobacteria	7.76	66.8
Burkholderia cenocepacia J2315	339	Bacteria	Proteobacteria	8.07	66.9
Burkholderia cenocepacia MC0-3	17929	Bacteria	Proteobacteria	7.9	66.6
Burkholderia mallei ATCC 23344	171	Bacteria	Proteobacteria	5.83	68.5
Burkholderia mallei NCTC 10229	13943	Bacteria	Proteobacteria	5.8	68.5

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Burkholderia mallei NCTC 10247	13946	Bacteria	Proteobacteria	5.9	68.5
Burkholderia mallei SAVP1	13947	Bacteria	Proteobacteria	5.2	68.4
Burkholderia multivorans ATCC 17616 JGI	17407	Bacteria	Proteobacteria	6.99	66.7
Burkholderia multivorans ATCC 17616 Tohoku	19401	Bacteria	Proteobacteria	6.99	66.7
Burkholderia phymatum STM815	17409	Bacteria	Proteobacteria	8.7	62.3
Burkholderia phytofirmans PsJN	17463	Bacteria	Proteobacteria	8.22	62.3
Burkholderia pseudomallei 1106a	16182	Bacteria	Proteobacteria	7.1	68.3
Burkholderia pseudomallei 1710b	13954	Bacteria	Proteobacteria	7.31	68
Burkholderia pseudomallei 668	13953	Bacteria	Proteobacteria	7	68.3
Burkholderia pseudomallei K96243	178	Bacteria	Proteobacteria	7.3	68.1
Burkholderia sp. 383	10695	Bacteria	Proteobacteria	8.69	66.3
Burkholderia thailandensis E264	10774	Bacteria	Proteobacteria	6.72	67.6
Burkholderia vietnamiensis G4	10696	Bacteria	Proteobacteria	8.4	65.7
Burkholderia xenovorans LB400	254	Bacteria	Proteobacteria	9.8	62.6
Campylobacter concisus 13826	17159	Bacteria	Proteobacteria	2.15	39.3
Campylobacter curvus 525.92	17161	Bacteria	Proteobacteria	2	44.5
Campylobacter fetus subsp. fetus 82-40	16293	Bacteria	Proteobacteria	1.8	33.3
Campylobacter hominis ATCC BAA-381	20083	Bacteria	Proteobacteria	1.7	31.7
Campylobacter jejuni RM1221	303	Bacteria	Proteobacteria	1.8	30.3
Campylobacter jejuni subsp. doylei 269.97	17163	Bacteria	Proteobacteria	1.8	30.6
Campylobacter jejuni subsp. jejuni 81116	17953	Bacteria	Proteobacteria	1.6	30.5
Campylobacter jejuni subsp. jejuni 81-176	16135	Bacteria	Proteobacteria	1.68	30.5
Campylobacter jejuni subsp. jejuni NCTC 11168	8	Bacteria	Proteobacteria	1.6	30.5
Candidatus Blochmannia floridanus	443	Bacteria	Proteobacteria	0.71	27.4
Candidatus Blochmannia pennsylvanicus str. BPEN	13875	Bacteria	Proteobacteria	0.79	29.6
Candidatus Carsonella ruddii PV	17977	Bacteria	Proteobacteria	0.16	16.6
*Candidatus Pelagibacter ubique HTCC1062	13989	Bacteria	Proteobacteria	1.3	29.7
Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)	16841	Bacteria	Proteobacteria	1.2	34
Candidatus Vesicomyosocius okutanii HA	18267	Bacteria	Proteobacteria	1	31.6
*Caulobacter crescentus CB15	298	Bacteria	Proteobacteria	4	67.2
Caulobacter sp. K31	16306	Bacteria	Proteobacteria	5.91	67.3
Cellvibrio japonicus Ueda107	28329	Bacteria	Proteobacteria	4.6	52
Chromobacterium violaceum ATCC 12472	444	Bacteria	Proteobacteria	4.8	64.8
Chromohalobacter salexigens DSM 3043	12636	Bacteria	Proteobacteria	3.7	63.9
Citrobacter koseri ATCC BAA-895	12716	Bacteria	Proteobacteria	4.71	53.8
Colwellia psychrerythraea 34H	275	Bacteria	Proteobacteria	5.37	38
Coxiella burnetii CbuG_Q212	19137	Bacteria	Proteobacteria	2	42.6

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Coxiella burnetii CbuK_Q154	19139	Bacteria	Proteobacteria	2.14	42.6
Coxiella burnetii Dugway 5J108-111	16721	Bacteria	Proteobacteria	2.25	42.4
Coxiella burnetii RSA 331	16791	Bacteria	Proteobacteria	2.04	42.7
Coxiella burnetii RSA 493	41	Bacteria	Proteobacteria	2.04	42.6
Cupriavidus taiwanensis	15733	Bacteria	Proteobacteria	6.46	67
Dechloromonas aromatica RCB	9635	Bacteria	Proteobacteria	4.5	59.2
Delftia acidovorans SPH-1	17413	Bacteria	Proteobacteria	6.8	66.5
Desulfococcus oleovorans Hxd3	18007	Bacteria	Proteobacteria	3.9	56.2
Desulfotalea psychrophila LSv54	12751	Bacteria	Proteobacteria	3.64	46.6
Desulfovibrio desulfuricans subsp. desulfuricans str. G20	329	Bacteria	Proteobacteria	3.73	57.8
Desulfovibrio vulgaris DP4	17227	Bacteria	Proteobacteria	3.7	63.2
Desulfovibrio vulgaris str. Hildenborough	51	Bacteria	Proteobacteria	3.8	63.3
Dichelobacter nodosus VCS1703A	50	Bacteria	Proteobacteria	1.4	44.4
Dinoroseobacter shibae DFL 12	17417	Bacteria	Proteobacteria	4.43	65.5
*Ehrlichia canis str. Jake	10694	Bacteria	Proteobacteria	1.3	29
*Ehrlichia chaffeensis str. Arkansas	325	Bacteria	Proteobacteria	1.18	30.1
*Ehrlichia ruminantium str. Gardel	13356	Bacteria	Proteobacteria	1.5	27.5
*Ehrlichia ruminantium str. Welgevonden v1	9614	Bacteria	Proteobacteria	1.5	27.5
*Ehrlichia ruminantium str. Welgevonden v2	13355	Bacteria	Proteobacteria	1.51	27.5
Enterobacter sakazakii ATCC BAA-894	12720	Bacteria	Proteobacteria	4.56	56.7
Enterobacter sp. 638	17461	Bacteria	Proteobacteria	4.66	52.9
Erwinia tasmaniensis Et1/99	20585	Bacteria	Proteobacteria	4.08	53.4
*Erythrobacter litoralis HTCC2594	13480	Bacteria	Proteobacteria	3.05	63.1
Escherichia coli 536	16235	Bacteria	Proteobacteria	4.9	50.5
Escherichia coli APEC O1	16718	Bacteria	Proteobacteria	5.51	50.3
Escherichia coli ATCC 8739	18083	Bacteria	Proteobacteria	4.7	50.9
Escherichia coli CFT073	313	Bacteria	Proteobacteria	5.2	50.5
Escherichia coli E24377A	13960	Bacteria	Proteobacteria	5.27	50.6
Escherichia coli HS	13959	Bacteria	Proteobacteria	4.6	50.8
Escherichia coli O157:H7 EDL933	259	Bacteria	Proteobacteria	5.59	50.3
Escherichia coli O157:H7 str. EC4115	27739	Bacteria	Proteobacteria	5.73	50.4
Escherichia coli O157:H7 str. Sakai	226	Bacteria	Proteobacteria	5.6	50.5
Escherichia coli SE11	18057	Bacteria	Proteobacteria	5.17	50.7
Escherichia coli SMS-3-5	19469	Bacteria	Proteobacteria	5.25	50.5
Escherichia coli str. K-12 substr. DH10B	20079	Bacteria	Proteobacteria	4.7	50.8
Escherichia coli str. K-12 substr. MG1655	225	Bacteria	Proteobacteria	4.6	50.8
Escherichia coli str. K-12 substr. W3110	16351	Bacteria	Proteobacteria	* 4.6	50.8
Escherichia coli UTI89	16259	Bacteria	Proteobacteria	5.21	50.6

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Francisella novicida U112	16088	Bacteria	Proteobacteria	1.9	32.5
Francisella philomiragia subsp. philomiragia ATCC 25017	27853	Bacteria	Proteobacteria	2	32.6
Francisella tularensis subsp. holarctica	16421	Bacteria	Proteobacteria	1.9	32.2
Francisella tularensis subsp. holarctica FTNF002-00	20197	Bacteria	Proteobacteria	1.9	32.2
Francisella tularensis subsp. holarctica OSU18	17265	Bacteria	Proteobacteria	1.9	32.2
Francisella tularensis subsp. mediasiatica FSC147	19571	Bacteria	Proteobacteria	1.9	32.3
Francisella tularensis subsp. tularensis FSC198	17375	Bacteria	Proteobacteria	1.9	32.3
Francisella tularensis subsp. tularensis SCHU S4	9	Bacteria	Proteobacteria	1.9	32.3
Francisella tularensis subsp. tularensis WY96-3418	18459	Bacteria	Proteobacteria	1.9	32.3
Geobacter bemidjiensis Bem	17707	Bacteria	Proteobacteria	4.6	60.3
Geobacter lovleyi SZ	17423	Bacteria	Proteobacteria	3.98	54.7
Geobacter metallireducens GS-15	177	Bacteria	Proteobacteria	4.01	59.5
Geobacter sulfurreducens PCA	192	Bacteria	Proteobacteria	3.8	60.9
Geobacter uraniireducens Rf4	15768	Bacteria	Proteobacteria	5.1	54.2
Gluconacetobacter diazotrophicus PAl 5	377	Bacteria	Proteobacteria	3.96	66.3
*Gluconobacter oxydans 621H	13325	Bacteria	Proteobacteria	2.92	60.8
*Granulibacter bethesdensis CGDNIH1	17111	Bacteria	Proteobacteria	2.7	59.1
Haemophilus ducreyi 35000HP	38	Bacteria	Proteobacteria	1.7	38.2
Haemophilus influenzae 86-028NP	11752	Bacteria	Proteobacteria	1.9	38.2
Haemophilus influenzae PittEE	16400	Bacteria	Proteobacteria	1.8	38
Haemophilus influenzae PittGG	16401	Bacteria	Proteobacteria	1.9	38
Haemophilus influenzae Rd KW20	219	Bacteria	Proteobacteria	1.8	38.1
Haemophilus somnus 129PT	322	Bacteria	Proteobacteria	2.01	37.2
Haemophilus somnus 2336	388	Bacteria	Proteobacteria	2.3	37.4
Hahella chejuensis KCTC 2396	16064	Bacteria	Proteobacteria	7.22	53.9
Halorhodospira halophila SL1	15767	Bacteria	Proteobacteria	2.7	68
Helicobacter acinonychis str. Sheeba	17251	Bacteria	Proteobacteria	1.6	38.2
Helicobacter hepaticus ATCC 51449	185	Bacteria	Proteobacteria	1.8	35.9
Helicobacter pylori 26695	233	Bacteria	Proteobacteria	1.67	38.9
Helicobacter pylori G27	31341	Bacteria	Proteobacteria	1.71	38.9
Helicobacter pylori HPAG1	16183	Bacteria	Proteobacteria	1.61	39.1
Helicobacter pylori J99	234	Bacteria	Proteobacteria	1.6	39.2
Helicobacter pylori P12	32291	Bacteria	Proteobacteria	1.71	38.8
Helicobacter pylori Shi470	29045	Bacteria	Proteobacteria	1.6	38.9
Herminiimonas arsenicoxydans	13467	Bacteria	Proteobacteria	3.4	54.3
*Hyphomonas neptunium ATCC 15444	15721	Bacteria	Proteobacteria	3.71	61.9

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Idiomarina loihiensis L2TR	10790	Bacteria	Proteobacteria	2.84	47
*Jannaschia sp. CCS1	12733	Bacteria	Proteobacteria	4.39	62.2
Janthinobacterium sp. Marseille	16549	Bacteria	Proteobacteria	4.1	54.2
Klebsiella pneumoniae 342	28471	Bacteria	Proteobacteria	5.88	56.9
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	31	Bacteria	Proteobacteria	5.69	57.1
Lawsonia intracellularis PHE/MN1-00	183	Bacteria	Proteobacteria	1.76	33.1
Legionella pneumophila str. Corby	17491	Bacteria	Proteobacteria	3.6	38.5
Legionella pneumophila str. Lens	13126	Bacteria	Proteobacteria	3.41	38.4
Legionella pneumophila str. Paris	13127	Bacteria	Proteobacteria	3.64	38.3
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	22	Bacteria	Proteobacteria	3.4	38.3
Leptothrix cholodnii SP-6	20039	Bacteria	Proteobacteria	4.9	68.9
Magnetococcus sp. MC-1	262	Bacteria	Proteobacteria	4.7	54.2
*Magnetospirillum magneticum AMB-1	16217	Bacteria	Proteobacteria	5	65.1
Mannheimia succiniciproducens MBEL55E	13068	Bacteria	Proteobacteria	2.3	42.5
*Maricaulis maris MCS10	17333	Bacteria	Proteobacteria	3.37	62.7
Marinobacter aquaeolei VT8	13239	Bacteria	Proteobacteria	4.75	56.9
Marinomonas sp. MWYL1	17445	Bacteria	Proteobacteria	5.1	42.6
*Mesorhizobium loti MAFF303099	18	Bacteria	Proteobacteria	7.6	62.5
*Mesorhizobium sp. BNC1	10690	Bacteria	Proteobacteria	4.94	61.1
Methylibium petroleiphilum PM1	10789	Bacteria	Proteobacteria	4.6	68.8
Methylobacillus flagellatus KT	10647	Bacteria	Proteobacteria	3	55.7
Methylobacterium extorquens PA1	18637	Bacteria	Proteobacteria	5.5	68.2
Methylobacterium populi BJ001	19559	Bacteria	Proteobacteria	5.85	69.4
Methylobacterium radiotolerans JCM 2831	18817	Bacteria	Proteobacteria	6.92	71
Methylobacterium sp. 4-46	18809	Bacteria	Proteobacteria	7.78	71.5
Methylococcus capsulatus str. Bath	21	Bacteria	Proteobacteria	3.3	63.6
Myxococcus xanthus DK 1622	1421	Bacteria	Proteobacteria	9.1	68.9
Neisseria gonorrhoeae FA 1090	23	Bacteria	Proteobacteria	2.15	52.7
Neisseria gonorrhoeae NCCP11945	29335	Bacteria	Proteobacteria	2.2	52.4
Neisseria meningitidis 053442	16393	Bacteria	Proteobacteria	2.2	51.7
Neisseria meningitidis FAM18	255	Bacteria	Proteobacteria	2.2	51.6
Neisseria meningitidis MC58	251	Bacteria	Proteobacteria	2.3	51.5
Neisseria meningitidis Z2491	252	Bacteria	Proteobacteria	2.2	51.8
*Neorickettsia sennetsu str. Miyayama	357	Bacteria	Proteobacteria	0.86	41.1
Nitratiruptor sp. SB155-2	18963	Bacteria	Proteobacteria	1.9	39.7
*Nitrobacter hamburgensis X14	13473	Bacteria	Proteobacteria	5.01	61.6
*Nitrobacter winogradskyi Nb-255	13474	Bacteria	Proteobacteria	3.4	62
Nitrosococcus oceani ATCC 19707	13993	Bacteria	Proteobacteria	3.54	50.3

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Nitrosomonas europaea ATCC 19718	52	Bacteria	Proteobacteria	2.81	50.7
Nitrosomonas eutropha C91	13913	Bacteria	Proteobacteria	2.82	48.5
Nitrosospira multiformis ATCC 25196	13912	Bacteria	Proteobacteria	3.25	53.9
*Novosphingobium aromaticivorans DSM 12444	204	Bacteria	Proteobacteria	4.23	65.1
Ochrobactrum anthropi ATCC 49188	19485	Bacteria	Proteobacteria	5.22	56.1
Oligotropha carboxidovorans OM5	28805	Bacteria	Proteobacteria	3.7	62.4
Orientia tsutsugamushi str. Boryong	16180	Bacteria	Proteobacteria	2.1	30.5
Orientia tsutsugamushi str. Ikeda	18983	Bacteria	Proteobacteria	2	30.5
*Paracoccus denitrificans PD1222	13020	Bacteria	Proteobacteria	5.25	66.8
Parvibaculum lavamentivorans DS-1	17639	Bacteria	Proteobacteria	3.9	62.3
Pasteurella multocida subsp. multocida str. Pm70	39	Bacteria	Proteobacteria	2.26	40.4
Pectobacterium atrosepticum SCRI1043	350	Bacteria	Proteobacteria	5.06	51
Pelobacter carbinolicus DSM 2380	13337	Bacteria	Proteobacteria	3.7	55.1
Pelobacter propionicus DSM 2379	13384	Bacteria	Proteobacteria	4.23	58.5
Phenylobacterium zucineum HLK1	19931	Bacteria	Proteobacteria	4.38	71.1
Photobacterium profundum SS9	13128	Bacteria	Proteobacteria	6.38	41.7
Photorhabdus luminescens subsp. laumondii TTO1	9605	Bacteria	Proteobacteria	5.69	42.8
Polaromonas naphthalenivorans CJ2	13418	Bacteria	Proteobacteria	5.35	61.7
Polaromonas sp. JS666	13121	Bacteria	Proteobacteria	5.9	62
Polynucleobacter necessarius subsp. asymbioticus QLW-P1DMWA-1	16679	Bacteria	Proteobacteria	2.2	44.8
Polynucleobacter necessarius subsp. necessarius STIR1	19991	Bacteria	Proteobacteria	1.6	45.6
Proteus mirabilis HI4320	12624	Bacteria	Proteobacteria	4.14	38.9
Pseudoalteromonas atlantica T6c	13454	Bacteria	Proteobacteria	5.19	44.6
Pseudoalteromonas haloplanktis TAC125	15713	Bacteria	Proteobacteria	3.84	40.1
Pseudomonas aeruginosa PA7	16720	Bacteria	Proteobacteria	6.6	66.4
Pseudomonas aeruginosa PAO1	331	Bacteria	Proteobacteria	6.3	66.6
Pseudomonas aeruginosa UCBPP-PA14	386	Bacteria	Proteobacteria	6.5	66.3
Pseudomonas entomophila L48	16800	Bacteria	Proteobacteria	5.9	64.2
Pseudomonas fluorescens Pf0-1	12	Bacteria	Proteobacteria	6.4	60.5
Pseudomonas fluorescens Pf-5	327	Bacteria	Proteobacteria	7.1	63.3
Pseudomonas mendocina ymp	17457	Bacteria	Proteobacteria	5.1	64.7
Pseudomonas putida F1	13909	Bacteria	Proteobacteria	6	61.9
Pseudomonas putida GB-1	17629	Bacteria	Proteobacteria	6.1	61.9
Pseudomonas putida KT2440	267	Bacteria	Proteobacteria	6.18	61.5
Pseudomonas putida W619	17053	Bacteria	Proteobacteria	5.8	61.4
Pseudomonas stutzeri A1501	16817	Bacteria	Proteobacteria	4.6	63.9
Pseudomonas syringae pv. phaseolicola 1448A	12416	Bacteria	Proteobacteria	6.08	57.9

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Pseudomonas syringae pv. syringae B728a	323	Bacteria	Proteobacteria	6.1	59.2
Pseudomonas syringae pv. tomato str. DC3000	359	Bacteria	Proteobacteria	6.54	58.3
Psychrobacter arcticus 273-4	9633	Bacteria	Proteobacteria	2.65	42.8
Psychrobacter cryohalolentis K5	13920	Bacteria	Proteobacteria	3.1	42.2
Psychrobacter sp. PRwf-1	15759	Bacteria	Proteobacteria	3.02	44.8
Psychromonas ingrahamii 37	16187	Bacteria	Proteobacteria	4.6	40.1
Ralstonia eutropha H16	13603	Bacteria	Proteobacteria	7.45	66.3
Ralstonia eutropha JMP134	10646	Bacteria	Proteobacteria	7.26	64.4
Ralstonia metallidurans CH34	250	Bacteria	Proteobacteria	6.91	63.5
Ralstonia pickettii 12J	17631	Bacteria	Proteobacteria	5.28	63.6
Ralstonia solanacearum GMI1000	13	Bacteria	Proteobacteria	5.8	67
*Rhizobium etli CFN 42	13932	Bacteria	Proteobacteria	6.53	61
Rhizobium etli CIAT 652	28021	Bacteria	Proteobacteria	6.44	61.3
Rhizobium leguminosarum bv. trifolii WSM2304	20179	Bacteria	Proteobacteria	6.87	61.2
*Rhizobium leguminosarum bv. viciae 3841	344	Bacteria	Proteobacteria	7.79	55
*Rhodobacter sphaeroides 2.4.1	56	Bacteria	Proteobacteria	4.61	68.8
Rhodobacter sphaeroides ATCC 17025	15755	Bacteria	Proteobacteria	4.54	68.2
*Rhodobacter sphaeroides ATCC 17029	15754	Bacteria	Proteobacteria	4.42	69
Rhodoferax ferrireducens T118	13908	Bacteria	Proteobacteria	4.97	59.6
*Rhodopseudomonas palustris BisA53	15751	Bacteria	Proteobacteria	5.51	64.4
*Rhodopseudomonas palustris BisB18	15750	Bacteria	Proteobacteria	5.51	65
*Rhodopseudomonas palustris BisB5	15749	Bacteria	Proteobacteria	4.89	64.8
*Rhodopseudomonas palustris CGA009	57	Bacteria	Proteobacteria	5.51	65
*Rhodopseudomonas palustris HaA2	15747	Bacteria	Proteobacteria	5.33	66
Rhodopseudomonas palustris TIE-1	20167	Bacteria	Proteobacteria	5.7	64.9
Rhodospirillum centenum SW	18307	Bacteria	Proteobacteria	4.4	70.5
*Rhodospirillum rubrum ATCC 11170	58	Bacteria	Proteobacteria	4.41	65.4
Rickettsia akari str. Hartford	12953	Bacteria	Proteobacteria	1.2	32.3
Rickettsia bellii OSU 85-389	17237	Bacteria	Proteobacteria	1.5	31.6
*Rickettsia bellii RML369-C	13996	Bacteria	Proteobacteria	1.52	31.6
Rickettsia canadensis str. McKiel	12952	Bacteria	Proteobacteria	1.2	31.1
*Rickettsia conorii str. Malish 7	42	Bacteria	Proteobacteria	1.3	32.4
*Rickettsia felis URRWXCal2	13884	Bacteria	Proteobacteria	1.59	32.5
Rickettsia massiliae MTU5	18271	Bacteria	Proteobacteria	1.41	32.5
*Rickettsia prowazekii str. Madrid E	43	Bacteria	Proteobacteria	1.11	29
Rickettsia rickettsii str. Iowa	19943	Bacteria	Proteobacteria	1.3	32.4
Rickettsia rickettsii str. Sheila Smith	9636	Bacteria	Proteobacteria	1.3	32.5
*Rickettsia typhi str. Wilmington	10679	Bacteria	Proteobacteria	1.11	28.9

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
*Roseobacter denitrificans OCh 114	16426	Bacteria	Proteobacteria	4.3	58.9
Saccharophagus degradans 2-40	316	Bacteria	Proteobacteria	5.1	45.8
Salmonella enterica subsp. arizonae serovar 62:z4,z23:--	13030	Bacteria	Proteobacteria	4.6	51.4
Salmonella enterica subsp. enterica serovar Agona str. SL483	20063	Bacteria	Proteobacteria	4.84	52
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	9618	Bacteria	Proteobacteria	4.99	52.1
Salmonella enterica subsp. enterica serovar Dublin str. CT_02021853	19467	Bacteria	Proteobacteria	4.88	52.1
Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	30687	Bacteria	Proteobacteria	4.7	52.2
Salmonella enterica subsp. enterica serovar Gallinarum str. 287/91	30689	Bacteria	Proteobacteria	4.7	52.2
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476	20045	Bacteria	Proteobacteria	4.99	52.1
Salmonella enterica subsp. enterica serovar Newport str. SL254	18747	Bacteria	Proteobacteria	4.98	52.2
Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601	30943	Bacteria	Proteobacteria	4.6	52.2
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150	13086	Bacteria	Proteobacteria	4.6	52.2
Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7	27803	Bacteria	Proteobacteria	4.9	52.1
Salmonella enterica subsp. enterica serovar Schwarzengrund str. CVM19633	19459	Bacteria	Proteobacteria	4.81	52.2
Salmonella enterica subsp. enterica serovar Typhi str. CT18	236	Bacteria	Proteobacteria	5.13	51.9
Salmonella enterica subsp. enterica serovar Typhi str. Ty2	371	Bacteria	Proteobacteria	4.8	52.1
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	241	Bacteria	Proteobacteria	4.99	52.2
Serratia proteamaculans 568	17459	Bacteria	Proteobacteria	5.45	55
Shewanella amazonensis SB2B	13385	Bacteria	Proteobacteria	4.3	53.6
Shewanella baltica OS155	13386	Bacteria	Proteobacteria	5.32	46.2
Shewanella baltica OS185	17643	Bacteria	Proteobacteria	5.28	46.3
Shewanella baltica OS195	13389	Bacteria	Proteobacteria	5.5	46.2
Shewanella denitrificans OS217	13390	Bacteria	Proteobacteria	4.55	45.1
Shewanella frigidimarina NCIMB 400	13391	Bacteria	Proteobacteria	4.85	41.6
Shewanella halifaxensis HAW-EB4	20241	Bacteria	Proteobacteria	5.2	44.6
Shewanella loihica PV-4	13906	Bacteria	Proteobacteria	4.6	53.7
Shewanella oneidensis MR-1	335	Bacteria	Proteobacteria	5.16	45.9
Shewanella pealeana ATCC 700345	17415	Bacteria	Proteobacteria	5.2	44.7
Shewanella piezotolerans WP3	17675	Bacteria	Proteobacteria	5.4	43.3
Shewanella putrefaciens CN-32	13393	Bacteria	Proteobacteria	4.7	44.5
Shewanella sediminis HAW-EB3	18789	Bacteria	Proteobacteria	5.5	46.1
Shewanella sp. ANA-3	13905	Bacteria	Proteobacteria	5.28	47.9
Shewanella sp. MR-4	13904	Bacteria	Proteobacteria	4.71	47.9

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Shewanella sp. MR-7	13903	Bacteria	Proteobacteria	4.8	47.9
Shewanella sp. W3-18-1	13902	Bacteria	Proteobacteria	4.7	44.6
Shewanella woodyi ATCC 51908	17455	Bacteria	Proteobacteria	5.9	43.7
Shigella boydii CDC 3083-94	15637	Bacteria	Proteobacteria	4.86	51
Shigella boydii Sb227	13146	Bacteria	Proteobacteria	4.63	51.1
Shigella dysenteriae Sd197	13145	Bacteria	Proteobacteria	4.59	51
Shigella flexneri 2a str. 2457T	408	Bacteria	Proteobacteria	4.6	50.9
Shigella flexneri 2a str. 301	310	Bacteria	Proteobacteria	4.82	50.7
Shigella flexneri 5 str. 8401	16375	Bacteria	Proteobacteria	4.6	50.9
Shigella sonnei Ss046	13151	Bacteria	Proteobacteria	5.03	50.8
*Silicibacter pomeroyi DSS-3	281	Bacteria	Proteobacteria	4.59	64.1
*Silicibacter sp. TM1040	13040	Bacteria	Proteobacteria	4.15	60.1
Sinorhizobium medicae WSM419	16304	Bacteria	Proteobacteria	6.82	61.1
*Sinorhizobium meliloti 1021	19	Bacteria	Proteobacteria	6.8	62.2
Sodalis glossinidius str. morsitans	16309	Bacteria	Proteobacteria	4.29	54.5
Sorangium cellulosum So ce 56	28111	Bacteria	Proteobacteria	13	71.4
Sphingomonas wittichii RW1	17343	Bacteria	Proteobacteria	5.93	67.9
*Sphingopyxis alaskensis RB2256	13907	Bacteria	Proteobacteria	3.37	65.5
Stenotrophomonas maltophilia K279a	30351	Bacteria	Proteobacteria	4.85	66.3
Stenotrophomonas maltophilia R551-3	17107	Bacteria	Proteobacteria	4.6	66.3
Sulfurimonas denitrificans DSM 1251	13019	Bacteria	Proteobacteria	2.2	34.5
Sulfurovum sp. NBC37-1	18965	Bacteria	Proteobacteria	2.6	43.9
Syntrophobacter fumaroxidans MPOB	13013	Bacteria	Proteobacteria	5	59.9
Syntrophus aciditrophicus SB	16258	Bacteria	Proteobacteria	3.2	51.5
Thiobacillus denitrificans ATCC 25259	13025	Bacteria	Proteobacteria	2.91	66.1
Thiomicrospira crunogena XCL-2	13018	Bacteria	Proteobacteria	2.4	43.1
Verminephrobacter eiseniae EF01-2	17187	Bacteria	Proteobacteria	5.63	65.2
Vibrio cholerae O1 biovar eltor str. N16961	36	Bacteria	Proteobacteria	4.03	47.5
Vibrio cholerae O395	15667	Bacteria	Proteobacteria	4.1	47.5
Vibrio fischeri ES114	12986	Bacteria	Proteobacteria	4.25	38.3
Vibrio fischeri MJ11	19393	Bacteria	Proteobacteria	4.48	38.2
Vibrio harveyi ATCC BAA-1116	19857	Bacteria	Proteobacteria	6.09	45.4
Vibrio parahaemolyticus RIMD 2210633	360	Bacteria	Proteobacteria	5.17	45.4
Vibrio vulnificus CMCP6	349	Bacteria	Proteobacteria	5.1	46.7
Vibrio vulnificus YJ016	1430	Bacteria	Proteobacteria	5.26	46.7
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	274	Bacteria	Proteobacteria	0.7	22.5
Wolbachia endosymbiont of Culex quinquefasciatus Pel	30313	Bacteria	Proteobacteria	1.5	34.2
*Wolbachia endosymbiont of Drosophila melanogaster	272	Bacteria	Proteobacteria	1.27	35.2

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
*Wolbachia endosymbiont strain TRS of Brugia malayi	12475	Bacteria	Proteobacteria	1.08	34.2
Wolinella succinogenes DSM 1740	445	Bacteria	Proteobacteria	2.1	48.5
Xanthobacter autotrophicus Py2	15756	Bacteria	Proteobacteria	5.62	67.3
Xanthomonas axonopodis pv. citri str. 306	297	Bacteria	Proteobacteria	5.27	64.7
Xanthomonas campestris pv. campestris str. 8004	15	Bacteria	Proteobacteria	5.15	65
Xanthomonas campestris pv. campestris str. ATCC 33913	296	Bacteria	Proteobacteria	5.08	65.1
Xanthomonas campestris pv. campestris str. B100	29801	Bacteria	Proteobacteria	5.1	65
Xanthomonas campestris pv. vesicatoria str. 85-10	13649	Bacteria	Proteobacteria	5.44	64.6
Xanthomonas oryzae pv. oryzae KACC10331	12931	Bacteria	Proteobacteria	4.9	63.7
Xanthomonas oryzae pv. oryzae MAFF 311018	16297	Bacteria	Proteobacteria	4.9	63.7
Xanthomonas oryzae pv. oryzae PXO99A	28127	Bacteria	Proteobacteria	5.2	63.6
Xylella fastidiosa 9a5c	271	Bacteria	Proteobacteria	2.73	52.6
Xylella fastidiosa M12	17823	Bacteria	Proteobacteria	2.5	51.9
Xylella fastidiosa M23	18457	Bacteria	Proteobacteria	2.54	51.7
Xylella fastidiosa Temecula1	285	Bacteria	Proteobacteria	2.52	51.8
Yersinia enterocolitica subsp. enterocolitica 8081	190	Bacteria	Proteobacteria	4.67	47.2
Yersinia pestis Angola	16067	Bacteria	Proteobacteria	4.68	47.6
Yersinia pestis Antiqua	16645	Bacteria	Proteobacteria	4.88	47.7
Yersinia pestis biovar Microtus str. 91001	10638	Bacteria	Proteobacteria	4.81	47.7
Yersinia pestis CO92	34	Bacteria	Proteobacteria	4.88	47.6
Yersinia pestis KIM	288	Bacteria	Proteobacteria	4.7	47.7
Yersinia pestis Nepal516	16646	Bacteria	Proteobacteria	4.61	47.6
Yersinia pestis Pestoides F	16700	Bacteria	Proteobacteria	4.71	47.7
Yersinia pseudotuberculosis IP 31758	16070	Bacteria	Proteobacteria	4.91	47.2
Yersinia pseudotuberculosis IP 32953	12950	Bacteria	Proteobacteria	4.8	47.6
Yersinia pseudotuberculosis PB1/+	28745	Bacteria	Proteobacteria	4.77	47.5
Yersinia pseudotuberculosis YPIII	28743	Bacteria	Proteobacteria	4.7	47.5
*Zymomonas mobilis subsp. mobilis ZM4	12354	Bacteria	Proteobacteria	2.06	46.3
Borrelia afzelii PKo	17057	Bacteria	Spirochaetes	* 1.24	27.8
Borrelia burgdorferi B31	3	Bacteria	Spirochaetes	1.52	28.2
Borrelia duttonii Ly	18231	Bacteria	Spirochaetes	1.57	28
Borrelia garinii PBi	12554	Bacteria	Spirochaetes	1.22	28
Borrelia hermsii DAH	29637	Bacteria	Spirochaetes	0.92	29.8
Borrelia recurrentis A1	18233	Bacteria	Spirochaetes	1.24	27.5
Borrelia turicatae 91E135	13597	Bacteria	Spirochaetes	0.92	29.1
Leptospira biflexa serovar Patoc strain Patoc	16153	Bacteria	Spirochaetes	3.95	38.9

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
1 (Ames)					
Leptospira biflexa serovar Patoc strain Patoc 1 (Paris)	20133	Bacteria	Spirochaetes	3.95	38.9
Leptospira borgpetersenii serovar Hardjo-bovis JB197	16148	Bacteria	Spirochaetes	3.9	40.2
Leptospira borgpetersenii serovar Hardjo-bovis L550	16146	Bacteria	Spirochaetes	3.92	40.2
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	10687	Bacteria	Spirochaetes	4.63	35
Leptospira interrogans serovar Lai str. 56601	293	Bacteria	Spirochaetes	4.66	35
Treponema denticola ATCC 35405	4	Bacteria	Spirochaetes	2.8	37.9
Treponema pallidum subsp. pallidum SS14	20067	Bacteria	Spirochaetes	1.1	52.8
Treponema pallidum subsp. pallidum str. Nichols	5	Bacteria	Spirochaetes	1.14	52.8
Acholeplasma laidlawii PG-8A	19259	Bacteria	Tenericutes	1.5	31.9
Aster yellows witches-broom phytoplasma AYWB	13478	Bacteria	Tenericutes	0.73	26.8
Candidatus Phytoplasma australiense	29469	Bacteria	Tenericutes	0.88	27.4
Candidatus Phytoplasma mali	25335	Bacteria	Tenericutes	0.6	21.4
Mesoplasma florum L1	10650	Bacteria	Tenericutes	0.79	27
Mycoplasma agalactiae PG2	16095	Bacteria	Tenericutes	0.88	29.7
Mycoplasma arthritidis 158L3-1	1422	Bacteria	Tenericutes	0.82	30.7
Mycoplasma capricolum subsp. capricolum ATCC 27343	16208	Bacteria	Tenericutes	1.01	23.8
Mycoplasma gallisepticum R	409	Bacteria	Tenericutes	1	31.5
Mycoplasma genitalium G37	97	Bacteria	Tenericutes	0.58	31.7
Mycoplasma hyopneumoniae 232	13120	Bacteria	Tenericutes	0.89	28.6
Mycoplasma hyopneumoniae 7448	10639	Bacteria	Tenericutes	0.92	28.5
Mycoplasma hyopneumoniae J	10675	Bacteria	Tenericutes	0.9	28.5
Mycoplasma mobile 163K	10697	Bacteria	Tenericutes	0.78	25
Mycoplasma mycoides subsp. mycoides SC str. PG1	10616	Bacteria	Tenericutes	1.2	24
Mycoplasma penetrans HF-2	176	Bacteria	Tenericutes	1.36	25.7
Mycoplasma pneumoniae M129	99	Bacteria	Tenericutes	0.82	40
Mycoplasma pulmonis UAB CTIP	100	Bacteria	Tenericutes	0.96	26.6
Mycoplasma synoviae 53	10676	Bacteria	Tenericutes	0.8	28.5
Onion yellows phytoplasma OY-M	9615	Bacteria	Tenericutes	0.86	27.7
Ureaplasma parvum serovar 3 str. ATCC 27815	19087	Bacteria	Tenericutes	0.75	25.5
Ureaplasma parvum serovar 3 str. ATCC 700970	101	Bacteria	Tenericutes	0.75	25.5
Ureaplasma urealyticum serovar 10 str. ATCC 33699	20247	Bacteria	Tenericutes	0.87	25.8
Fervidobacterium nodosum Rt17-B1	16719	Bacteria	Thermotogae	1.9	35
Petrotoga mobilis SJ95	17679	Bacteria	Thermotogae	2.2	34.1

Organism Name	NCBI Project ID	Domain	Phylum	Genome Size (Mb)	G+C Content
Thermosipho melanesiensis BI429	17249	Bacteria	Thermotogae	1.9	31.4
Thermotoga lettingae TMO	15644	Bacteria	Thermotogae	2.1	38.7
Thermotoga maritima MSB8	111	Bacteria	Thermotogae	1.86	46.2
Thermotoga petrophila RKU-1	17089	Bacteria	Thermotogae	1.8	46.1
Thermotoga sp. RQ2	19543	Bacteria	Thermotogae	1.9	46.2
Akkermansia muciniphila ATCC BAA-835	20089	Bacteria	Verrucomicrobia	2.7	55.8
Methylacidiphilum infernorum V4	28995	Bacteria	Verrucomicrobia	2.3	45.5
Opitutus terrae PB90-1	19989	Bacteria	Verrucomicrobia	6	65.3