4-1-2011

COMPUTATIONAL GENOMIC SIGNATURES AND METAGENOMICS

Ozkan U. Nalbantoglu
University of Nebraska-Lincoln, ufuknalbantoglu@yahoo.com

COMPUTATIONAL GENOMIC SIGNATURES AND METAGENOMICS

by

Ozkan Ufuk Nalbantoglu

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Engineering (Electrical Engineering)

Under the Supervision of Professor Khalid Sayood

Lincoln, Nebraska

April, 2011

COMPUTATIONAL GENOMIC SIGNATURES AND METAGENOMICS

Ozkan Ufuk Nalbantoglu

University of Nebraska, 2011

Advisor: Khalid Sayood

Mathematical characterizations of biological sequences form one of the main elements of bioinformatics. In this work, a class of DNA sequence characterization, namely computational genomics signatures, which capture global features of these sequences is used to address emerging computational biology challenges. Because of the species specificity and pervasiveness of genome signatures, it is possible to use these signatures to characterize and identify a genome or a taxonomic unit using a short genome fragment from that source. However, the identification accuracy is generally poor when the sequence model and the sequence distance measure are not selected carefully. We show that the use of relative distance measures instead of absolute metrics makes it possible to obtain better detection accuracy. Furthermore, the use of relative metrics can create opportunities for using more complex models to develop genome signatures, which cannot be used efficiently when conventional distance measures are used.

Using a relative distance measure and a model based on the relative abundance of oligonucleotides in a genome fragment, a novel genome signature was defined. This signature was employed to address a class of metagenomics problems. The metagenomics approach enables sampling and sequencing of a microbial community without isolating and culturing single species. Determining the taxonomic classification of the bacterial species within the microbial community from the mixture of short DNA fragments is a difficult computational challenge. We present supervised and unsupervised algorithms for taxonomic classification of metagenomics data and demonstrate their effectiveness on simulated and real-world data. The supervised algorithm, RAIphy, classifies metagenome fragments of unknown origin by assigning them to the taxa, defined in a signature database of previously sequenced microbial genomes. The signatures in the database are updated iteratively during the classification process. Most metagenomics samples include unidentified species, thus they require clustering. Pseudo-assembly of fragments, followed by clustering of taxa is employed in the unsupervised setting. The signatures developed in this work are more specific-specific and pervasive than any signatures currently available in the literature, and demonstrate the potential and viability of using genome signatures to solve various metagenomics problems as well as other challenges in computational biology.

ACKNOWLEDGEMENTS

It has been a long journey, timely and mentally. During the journey I evolved from *mitos* to *logos*. I would like to thank my advisor Dr. Khalid Sayood, for his wisdom, guidence and I am grateful for his unbelievable patience with me. He has always been the wise one having the solutions of any sort of problems, during the long process. He has always given me the freedom to work on what I find interesting, and the has always given me the freedom to make mistakes. I feel lucky to adopt some of his point of view for approaching science and engineering problems. The concept of academic inheritance sounds more meaningful to me now, and I am glad to have such a valuable inheritance.

I feel fortunate to have my committee members, Dr. Sina Balkir, Dr. Stephen D. Scott, and Dr. Michael W. Hoffman. I appreciate Dr. Balkir's guidance, encouragement, and his continuous eagerness to provide helpful advices. I would like to thank Dr. Hoffman, who is the one that can always read between the lines. His concise advices has been to the point, that never failed to open my way. During a couple of computer science classes, Dr. Scott triggered my interest in some topics. These topics were important corner stones during the development of my dissertation. I also want to thank him for being such a great teacher, and also for putting a great effort in reading my dissertation and providing great feedback.

I would like to thank the department for the financial support, and all members of the department for being extremely helpful. I especially want to thank my colleagues, Dr. David Russell and Sam Way. Dave has positively changed my approach to the problems, and Sam, being a great programmer, has worked with me with a great harmony.

My thanks also go to my dear friends Marina Marshenkulova, Leyla Mas-

maliyeva, Luisa Adamyan, Farooq Shams, Karina Deifel, Saby Sarkar and Ami Sarkar who have accompanied me during this journey, and have been almost family to me. Last but not least, I would like to thank my family, who always supported me with my decisions.

My personal opinion is that any inspiration provided by this work to another researcher would be more valuable than the dissertation *per se*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Genomes can be viewed as linear strings of four bases, adenine (A), guanine (G), cytosine (C), and thymine (T). This enables the treatment of genome sequences as symbolic sequences and the characterization of these sequences using mathematical models. The mathematical models of genomic sequences of particular interest to us in this work is a class of models called genome signatures. Genome signatures are compact mathematical representations of DNA sequences. They characterize the sequences in a manner that emphasizes features specific to the organism from which the DNA was obtained. Examples of such signatures are parametric models that make use of statistics gathered from fragment of the DNA sequence. In the case of genome signatures the estimated parameters are unique to a species; therefore, genome signatures constitute species-specific characterization of DNA sequences.

A second attribute of genome signatures that make them a potentially significant tool for bioinformatics applications is the pervasiveness of the signature. By pervasiveness of a signature we mean that species specificity of the signature is preserved for any arbitrary genome fragment. According to this property, different genome fragments from the same genome have similar mathematical characterizations. Moreover, these characterizations are similar for varying fragment lengths. The two properties of species specificity and pervasiveness determine the strength of a genome signature. A strong signature, which is highly species-specific and pervasive can characterize a genome using only a small random part of the genome. For many bioinformatics applications, detection of the species of origin from small random genome fragments is required. Genome signatures are good candidates for such tasks.

In practice, the species specificity and pervasiveness of signatures is limited by many factors. Consider a mathematical characterization in the form of a parametric model where the model parameters are estimated using the statistics gathered from the genome fragments. While the statistics obtained from long genome fragments could provide good estimates, poor estimates due to insufficient statistics may be observed in the case of short sequences. Poor estimates of the signature parameters result in weak signatures which are not very efficient for distinguishing between various candidate organisms as the species of origin. This is a common problem, and most known genome signatures suffer from this problem.

Poor specificity and pervasiveness problems force researchers to use simple genome signatures that do not require large number of parameters to estimate. However, use of simple structures can also lead to poor characterization. This phenomenon is a major obstacle to the use of genome signatures in various bioinformatics applications. As an example, genome-signature based methods are widely used for long contigs in taxonomy assignment applications of metagenomics. However, a general trend is to employ database search methods for the assignment of short DNA sequences to their taxonomic origin, in spite of their computational burden, because genome-signature based methods mostly fail at this task.

1.1 Contributions of this Dissertation

Our fundamental observation is that species specificity and the pervasiveness of a genome signature do not only depend on the structure of the characterization, but also depend on how the distances/similarities between the signatures are measured. We claim that, by an appropriate selection of the distance metric, more information contained in a signature can be exploited. Conventional use of signatures mostly employs absolute distance/similarity metrics such as Euclidean metrics, correlation measures, etc. However, we show that when relative measures, such as model fitness or likelihood function calculations replace these absolute measures, it is possible to obtain better detection accuracy. Similarly, the use of relative metrics can create the opportunity for using more complex models, which cannot be used efficiently with conventional measures, as signatures.

Based on this observation we have developed signatures and similarity/difference measures that are superior to any combination currently available in the literature. As an application of the signature developed in this work, a supervised metagenome binning algorithm called *RAIphy* [1] is proposed. RAIphy outperforms all currently known compositional based binning programs for a broad range of fragment lengths. The performance of RAIphy is competitive with similarity-search methods although RAIphy has much lower computational complexity.

We have also considered the metagenome binning task in an unsupervised setting using the same principle of dependence between the signature and the similarity/difference metric. Unsupervised RAIphy is an algorithm that combines concepts from genome assembly and metagenome binning for unsupervised taxonomic grouping. Our tests show the superior performance for unsupervised RAIphy when compared to currently popular unsupervised metagenome binning methods.

The framework studied for the efficient use of genome signatures is promising for further applications of bioinformatics because it implies that genome signatures might obtain more information about a genome than previously understood. This opens up the potential for further applications.

1.2 Organization of this Dissertation

In the following chapter we introduce the concept of genome signatures and their historical development. In this chapter we focus mainly on signatures that are based on the frequency of occurrence of short oligonucleotides. In Chapter 3 we continue with our discussion of mathematical characterizations which could be used as genome signatures because of their properties of species specificity and pervasiveness but which are not directly based on the frequency of occurrence of short oligonucleotides. The use of relative distance/similarity metrics and their advantages are discussed in Chapter 4. We also introduce a novel metric called the *Relative Abundance Index* in this chapter. An introduction to metagenomics is provided in Chapter 5. Chapter 6 introduces the metagenome binning algorithm, RAIphy. An unsupervised version of RAIphy supported with a novel metagenome binning paradigm is presented in Chapter 7. Chapter 8 contains the summary and further research directions.

Chapter 2

Genome Signatures, Definition and Background

Since the discovery of the fact that the deoxyribonucleic acid (DNA) is the primary repository of genetic information, the understanding of the molecular evolution of biological sequences such as DNA, RNA and proteins has been invaluable for understanding the driving forces, trends and implications of the evolution of species. Development of statistical tools for analyzing biological sequences has been useful for capturing the effect of evolution on genomes. An important discovery in this direction is that the compositional features of a genome carry information about the evolutionary history of a species.

These compositional features carry specific signals which permit organisms to be distinguished on the basis of genus and species. This specificity can be interpreted to be the result of the adaptation of the species process to the environment. Observed environmental and structural parameters are some of the factors shaping DNA, RNA and protein compositions. Furthermore, physicochemical structural constraints and high level cellular machinery also shape the organization of biological sequences.

Along with providing a means for distinguishing between species, the species specificity can be used in a number of ways. The relative homogeneity of the compositional factors means that the species-specificity of these features exists throughout the genome. These two properties of species specificity and pervasiveness are major components of a genomic signature.

2.1 Definition of Computational Genomic Signatures

Characterizations of species specific features in biological sequences are often described by the term *signature.* The term genomic signature has been used homonymously corresponding to similar concepts, but to different properties. For instance, a species specific feature obtained from a genome is frequently used as a genome signature. Such a feature may be a short fragment of the genome unique to the organism. A sequence of around 20-25 bp in length has a low probability of appearing in all genomes. Therefore, those sequences are comprehensively searched for and labeled as barcodes belonging to specific taxonomic groups. A detection technology, such as microarray platforms [2] or PCR assays [3,4], can detect these barcodes resulting in the detection of the unknown organism. This barcoding methodology has been used for building catalogues of species and identification of birds [5], fishes [6] and amphibians [7] as well as a large set of other eukaryotes[8]. Similarly, barcoding using composition vectors gathered from rRNA sequences has also been used for similar purposes [9,10]. The genomic signature in this sense is located in a specific region of the genomes, and the knowledge of the entire genome or at least the location and sequence of that region is required for defining the genome signature.

Unlike previous genome signature definitions, computational genomic signatures utilize the relative homogeneity of genomes as well as the species specificity of DNA. A *computational genomic signature* is a species-specific mathematical structure that can be generated from an arbitrary genome fragment. That is to say, given a random fragment of any genome with sufficient length, one can generate the same (or similar) mathematical characterization for a given genome. The resulting structure is distinguishable from that obtained from the genome of a different organism. In order to introduce the distinguishability of signatures, a metric is also needed in the space where the signature is defined. That is:

$$d_{\mathcal{S}}(S(G_{X_i}), S(G_{X_j})) < d_{\mathcal{S}}(S(G_{X_i}), S(G_{Y_k})), \tag{2.1}$$

where G_{X_i} and G_{X_j} are random DNA sequences from the genome G_X and G_{Y_k} is a random DNA sequence from genome G_Y, and $i, j, k \in \mathbb{N}^+$. $S(.)$ is an operation over the domain of possible DNA sequences and the range of $S(.)$ exists in a metric signature space. The distances in this signature space are shown with the metric $d_{\mathcal{S}}(.,.)$. Ideally, the signature is embedded in any subsequence of a genome, that is $d_{\mathcal{S}}(S(G_{X_i}), S(G_{X_j})) = 0$. In practice, due to the heterogeneities introduced by functional constraints and random mutations/deletions/insertions, these intergenomic distances are generally non-zero. These intergenomic distances depend on both the feature extraction ability of the signature and the metric defined in the signature space. Two attributes determine the quality of a genome signature: *species specificity*, and *pervasiveness*. Genome signatures are pervasive, in that they appear throughout the genome, and species specific, in that they are different for different organisms.

2.2 Compositional Features as Genome Signatures

2.2.1 GC Content:

GC content, an early discovered compositional feature of genomes, is a popular characterization, which satisfies the genome signature definition. It measures the ratio of cytosine + guanine bases in a DNA sequence. The ratio of genomic GC content is biased accross the tree of life and ranges from 16.5% (Carsonella ruddii) to 75% (Anaeromyxobacter dehalogens) [11,12]. GC variation is also correlated with phylogenetic variation [13].

The variation of GC-content has been attributed to several factors. The difference in physicochemical character of the cytosine – guanine and adenine – thymine bonds results in varying reactions to different factors. Examples include cytosine and guanine forming 3 H-bonds between the strands in the double helix and being more resistant to denaturation [14], different reaction to reactive oxygen species damage [10], the availability and lower cost of A/T products, the preference of GC over AT in different respiratory behavior, growth temperatures and ecological conditions. Along with the selective perspective maintaining that GC bias is driven by selective pressures exerted by the environment, there is also a naturalist

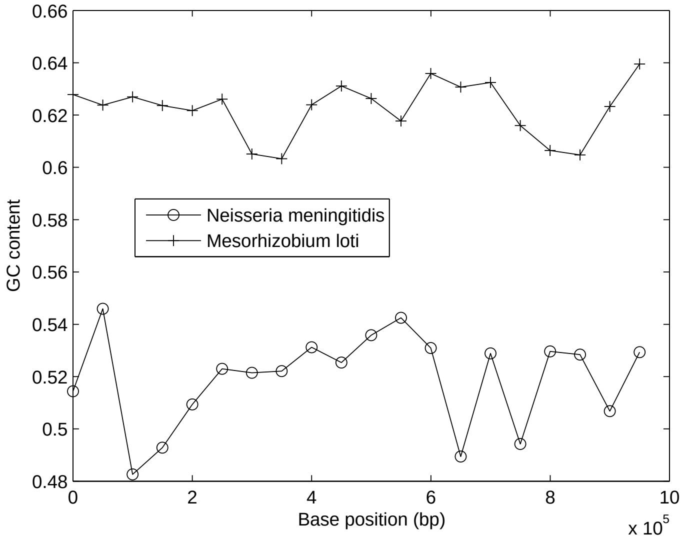

Figure 2.1: GC-content of randomly chosen 50 kb genomic fragments of *Neisseria meningitidis* and *Mesorhizobium loti.*

camp which claims that [15,16] the bias is not a result of selection but is due to a neutral mutational behavior. Because of their variation with varying environmental parameters, GC-content values appear to be species-specific. Moreover, as the bases are distributed throughout a genome in similar proportions, the GC content satisfies the pervasiveness attribute of a genome signature.

The different values of GC content for various species, and its relative conservation within a genome was noticed in the early 1960s [17]. We can observe the genome signature property of GC-content in randomly chosen 50 kb genomic fragments of *Neisseria meningitidis* and *Mesorhizobium loti* as shown in Figure 2.1. The GC-content is also the simplest form of signatures, since it has only one rational number parameter and the distance metric is simply the arithmetic difference of these values.

Fvalue	synonymous codon usage	Amino acid usage
Genus	969.55	708.65
Family	1016.85	818.15
Order	1186.3	875.54

Table 2.1: F-scores of one-way-ANOVA for amino acid usage and synonymous codon usage. Distribution of profiles at different clade levels are considered.

2.2.2 Amino acid content:

Amino acid content represents the relative frequencies of amino acids used in a protein or a proteome with a 20 dimensional vector. It involves the simplest feature at the proteome level, analogous to GC content at the genome level. Certain organisms prefer different amino acids in their proteins, resulting in a spectrum of typical amino acid usage of various taxa.

It has been suggested that the species specificity of amino acid usage is the outcome of certain evolutionary processes. Response to different environmental temperatures [18,19], economy of nutrient supply [20-22], susceptibility to oxidation and the resulting behavior under different respiratory regimes [23] are among the factors shaping the amino acid content.

The preference for certain amino acids is also fairly conserved throughout a genome. Because genes do not diverge significantly in the preference of amino acids they code, this preference is pervasive through the genome. This signature property was used by Sandberg et. al. for classification of proteins based on their amino acid content.

2.2.3 Synonymous codon usage:

Synonymous codon usage is generally represented by 64 dimensional vectors which reflect the relative frequency of each codon coding for an amino acid. In the early 1980s, it was noted that each species systematically prefers certain codons to code an amino acid; this phenomenon is true for most genes of an organism [25-27]. The proposition that synonymous codon usage is species specific is known as Grantham's genome hypothesis.

The variation of synonymous codon usage among the genes of an organism is frequently attributed to gene expression levels and the relative abundance of tRNA's in a cell [28,29]. Variation between genomes is more significant than intergenomic variation. Even though the usage of synonymous codons does not change the protein composition, it has also been linked to amino acid composition [30-33], protein structure [34-36], directional mutational biases [37-39], and mRNA secondary structure [40]. The direct relationship of synonymous codon usage to the environmental factors can be seen by the fact that synonymous codon usage carries signals revealing information about the thermal and respiratory behavior of an organism [41].

Following a similar statistical methodology used for amino acid usage, it was also shown that synonymous codon usage exhibits genome signature characteristics [24]. Table 2.1 shows one-way ANOVA test results based on F-scores for amino acid usage and synonymous codon usage. Each gene was represented by its amino acid/synonymous codon usage profile and analysis of variance is employed assuming each taxon as one group. The test is performed for the clade levels of genus, family and order. Higher F-scores imply a clearer separation of taxa in the vector spaces of the corresponding genome signatures.

2.3 Methods of Characterization Embedded in the Initial Work on DNA

In the 1960s, the first glimpses of genome signatures appeared as supplementary observations to the experiments designed for different purposes. Before the birth of computational biology, with the non-existence of molecular databases and *in silico* genome analysis, Kornberg and colleagues [42,43] conducted a series of studies using the replication factors from phage ΦX 174 and primer sets to synthesize DNA of viral, bacterial, plant and animal sources. The ingenious technique they used involved $5' - P^{32}$ labeled DNA to obtain the percentage of different dinucleotides. Their main motivation and thus the main observation was confirming Watson-Crick base pairing by comparing the reverse complement doublets in forward and reverse strands. Along with achieving their primary goal, they also found that the frequency of occurrence of dinucleotides did not follow a random model. That is, the frequency of occurrence of a dinucleotide pair XY was not equal to the product of the frequency of occurrence of each individual nucleotide X and Y. They also found that the dinucleotide frequencies obtained from different taxonomies such as mouse tumors, crab testis, bovine liver, as well as plants and viral DNA were distinguishable by dinucleotide frequencies. In particular, they found that the frequency of occurrence of the CpG dinucleotide fits a random model for bacteria, but it moves progressively away from a random model for echinoderms and vertebrates. Another important observation they reported was that the synthesized DNA sequences had the same doublet frequency characteristics with the primers used to synthesize these sequences for viral, bacterial and animal sources. These additional observations are actually indications of the species specificity and pervasiveness of doublet frequencies as genome signatures. Subak-Sharpe and colleagues [44,45] defined the term “general design of an organism” as the normalized frequency of occurrence (odds ratio) of dinucleotides, and they noted the similarity of the general design of several small mammalian viruses and their hosts [46].

2.4 Dinucleotide Odd-ratios as a Genome Signature

After the first indications of the existence of genomic signatures, it took almost 30 years to reconsider the concept. With the increasing availability of genomic sequences, Karlin and colleagues, in a sequence of papers [47-55], extended the work of Kornberg et al., and Subak-Sharpe et al.; and coined the term *genomic signature*. Initially, the odds ratio of dinucleotides (along with tri- and tetranucleotides) to measure the divergence of neighboring bases from expected distributions was introduced to observe the over- and underrepresentation of dinucleotides in genomes [47]:

$$\rho^*_{XY} = \frac{f^*(XY)}{f^*(X)f^*(Y)}. \tag{2.2}$$

Here $f^*(XY)$ stands for the frequency of the dinucleotide XY in the given fragment concatenated with its reverse strand. $f^*(X)$ and $f^*(Y)$ are the frequencies of the bases X and Y. This odd-ratio gives an overrepresentation or an underrepresentation measure for the all 16 dinucleotides. Note that $f^*(X)$ values are calculated using both strands. The frequencies without star superscripts are the frequencies calculated using one strand of the genome. Because of Watson-Crick pairing, $f^*(G) = f^*(C) = f(G + C)$; the same property applies for A and T. Initially, these measurements were used individually, and global properties of dinucleotide occurrence, such as the underrepresentation of AT in almost all taxonomies, underrepresentation of CG in vertebrates and mitochondrial DNA, and overrepresentation of homodimers, along with the corresponding evolutionary implications, were discussed. Karlin and Ladunga [49] examined the normalized frequency of occurrence of di-, tri- and tetra-nucleotides in various eukaryotic genomes. As a result of this study, they noted that the Euclidean distance of relative abundance profiles for closely related organisms were smaller than the distances calculated for phylogenetically distant organisms. Later on, a metric which took into account all 16 dinucleotide abundance values, the δ distance, was introduced [54]:

$$\delta^*(f, g) = 1/16 \sum_{XY} |\rho^*_{XY}(f) - \rho^*_{XY}(g)|. \tag{2.3}$$

Having defined two requirements for a genome signature, the signature and the distance metric in signature space, Karlin et al. investigated the species specificity and pervasiveness of that signature. It was seen that δ distance is very small within the same species, being only 2-3 times the distance found in random DNA. Another result was that within the genome, the distance is generally smaller than the intergenomic measurements. In fact, in some cases, the species specificity and pervasiveness of dinucleotide relative abundance ratio profiles are even visible to the naked eye without any metric definition. An example is shown in Figure 2.2

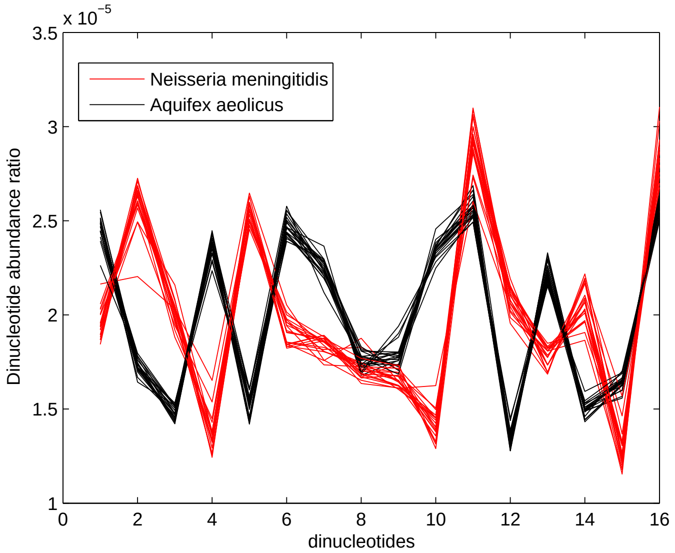

Figure 2.2: The dinucleotide odds-ration profiles for 20 random 50 kbp segments from *Neisseria meningitidis* and *aquifex aeolicus* genomes.

for 20 random 50 kbp segments from *Neisseria meningitidis* and *aquifex aeolicus* genomes. Clearly, the 50 kbp sections are distinguishable for these two genomes.

A fruitful series of applications followed this initial discovery of genome signatures. The dinucleotide abundance signature along with δ distance has been observed to be pervasive also in Eukaryotes for $\geqslant 50kbp$ genomic fragments. Moreover, according to their genome signature analysis, archea appeared to be an inconsistent clade having large signature distance between the members. Although the dinucleotide abundance profiles of nuclear DNA and mitochondrial DNA are significantly different than each other, it was found that the distances between the mitochondrial DNA are in parallel with the distances obtained from nuclear DNA segments. This result was considered as quantitative evidence for the coevolution of eukaryote cells and their mitochondria. Moreover, the mitochondria of mammals were reported as being very similar to each other, while animal and fungal mitochondria DNA were moderately similar and all very different than plant and protist mitochondrial sequences. With their genome signature studies on virus and bacterial plasmids, Karlin and colleagues found that both virus and plasmids resemble the structure of their hosts. Also among viral genomes, single stranded RNA viruses are found to be the species having the most obscure signatures which are close to random sequences. They attributed that random nature to the high mutation rate of single stranded RNA.

During their investigation of genomic signatures, Karlin and colleagues were not able to determine a clear relationship between environmental factors (e.g. habitat propensities, osmolarity tolerance, chemical conditions) and their genome signature. They mostly attributed the emergence of signatures to the structural properties of the DNA polymer such as dinucleotide stacking energies, curvature, chromosomal organization, DNA packaging, DNA replication, transcription, and repair mechanisms.

2.5 Chaos Game Representation

The history of genome signature discovery has evolved from two different biological sequence analysis camps. The first group contains the initial *in vivo* approaches investigating dinucleotide occurrence frequencies. In this approach, the over-and underabundance of nucleotide doublets accounts for species specificity and pervasiveness. Another branch of sequence analysis followed statistical mechanics approaches to analyze the genomic sequences, finally ending up with another form of genome signatures. Later on, the tight connection between those two concepts being instances of oligonucleotide composition was reported.

The attempts to represent genome sequences in other mathematical forms, in which a rich repertoire of analysis tools is available, has been of great interest to researchers. Some of these approaches have their roots in statistical mechanics. Representing the sequences as random walks [56-59] has revealed some features, such as the walks of DNA sequences resembling fractal behavior. Moreover, divergence from random sequences and exhibiting Markov-like behavior provided a basis for further investigation of compositional features. In 1990 Jeffrey [60] proposed a method he called the Chaos Game Representation (CGR) to visualize the genomic sequences. This was a method employed from nonlinear dynamics [61], as a two dimensional representation of symbolic sequences. According to this scheme, a symbolic sequence is scanned with a running window of length k, and with every step the observed k-mer is represented in a 2 dimensional iterated map. Simply, we can assume that from the left-top quadrant in clockwise direction each quadrant represents C, G, T, and A respectively in a square. The first base is placed in the corresponding quadrant, after that the quadrant is divided into 4 quadrants, and the same procedure is applied for the second base. Iteratively, the observed window finds its place in one of the 4^k squares, in k steps of iteration. Complex nonrandom symbolic sequences are observed to form fractal images with chaos game representation. Jeffrey observed this behavior in DNA sequences and concluded that DNA sequences were far from random.

An objection to chaos game representation of genomic sequences arose from Goldman, claiming that it reflects the short term correlations of DNA rather than capturing complex structures. He added the claim that the same images can be generated from mono-, di- and trinucleotide frequencies of DNA sequences. Indeed, he was able capture the “double scoop” character, an indication of scarcity

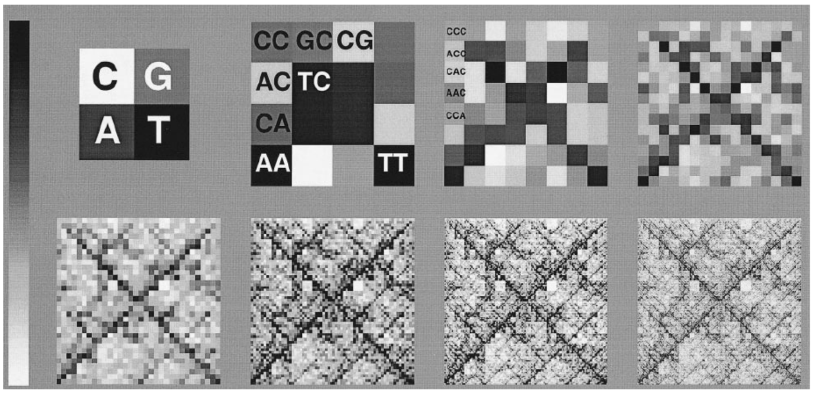

Figure 2.3: The generation of the CGR of *Archeoglobus Fulgidus* genome in 8 iterations. (figure taken from Deschevanne et al [62])

in the CG doublet, of CGR observed in vertebrate genomes and in vertebrate viruses. Goldman was right in claiming that CGR images do not capture complex structures but reflect the short term correlations, and he was wrong in claiming that those images do not provide superior information than that obtained from oligonucleotide frequencies up to trinucleotide or even the codon usage. In fact, CGR images contain the information of k-mers and not more than that. Since the correlations in DNA is longer than 3 base separation dependencies, CGR can provide better knowledge than codon usage. To see how CGR images exactly contain k-mer frequency information clearly we can follow this reading: The idea of this representation is the whole set of frequencies from mononucleotide frequencies to k-length word frequencies found in a given genomic sequence can be displayed in the form of a single image in which each pixel is associated with a specific word. The difference of this specific representation from a random arrangement of pixels in the image is that the generation of the image is a recursive process starting from 4 pixels for mononucleotides and splitting each pixel by 4 in every iteration for each word length expansion. This can be thought of as increasing the resolution of a quantized image. The grayscale value indicates the relative frequency of a word; darker values indicate greater relative frequency values. In Figure 2.3 the generation of the CGR of *Archeoglobus Fulgidus* genome is shown. The resulting images show certain characteristics as the word length increases. A human expert can comprehend the characteristics of a genome by analyzing the CGR. For instance, the lighter upper part indicates low G+C composition, diagonally oriented lines represent the abundance of purine and pyrimidine stretches. These diagonal lines can be seen in the Figure 2.3.

It was Deschevanne et al. [62], who discovered that CGR representation could also be used as a signature. With CGR images created from different organisms, it was clear that different organisms attain distinguishable CGR images. Moreover, the images obtained from random genomic fragments down to 1000 bp in length formed images resembling different variations of the same image to the human eye (Figure 2.4).

Heuristically the pervasiveness and species specificity of CGR images are visible. However, as signatures are mathematical structures there is a need for metrics to quantify the signature behavior as mentioned before. Euclidian distances between the CGR images, obtained by adding the squared pixel differences for the same pixel locations, were calculated, and the species specificity and pervasiveness were shown by computational experiments [63]. It is clear that the Euclidian distances of 2^k times 2^k CGR images correspond to the vector distances of k-mer frequencies in the composition space. Therefore, the discussion reduces to the fact

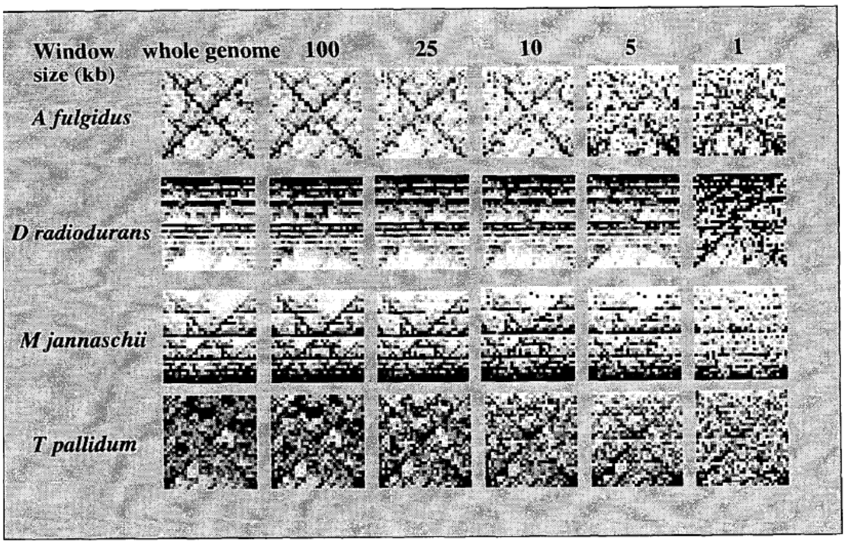

Figure 2.4: CGR images for A fulgidus, D radiodurans, M jannaschii, and T pallidum for varying fragment length. (figure taken from Deschevanne et al [62])

that oligonucleotide frequencies are genome signatures. A close relationship between the dinucleotide abundance ratio signatures and GCR images was noticed by Wang et al. [64], and it was concluded that as the information of dinucleotide abundance profiles are already embedded in CGR images and they belong to a spectrum of genomic signatures. These results imply that, dinucleotide abundance ratios, CGR and oligonucleotide frequencies are computational genomic signatures of the same class.

2.6 A Unified Framework of Genome Signatures: Functions of Oligonucleotide Occurrence

It is possible to define a general compositional feature from which the genome signatures defined above can be deduced. A general scheme serving this purpose is the frequency of oligonucleotide occurrence in a DNA fragment. GC content, synonymous codon usage, and amino acid content can be approximately expressed as functions of oligonucleotide frequency profiles. Moreover, genome signatures defined on dinucleotide abundance ratios and CGR images are functions of oligonucleotide frequencies.

Given a oligonucleotide frequency vector of a DNA sequence (with the oligonucleotide length of k), the GC content of this sequence can be obtained by summing up the first $2^{(2k-1)}$ components of this vector. As an example, we can take a random 100 kbp fragment of *e. coli* genome and look at the relative dinucleotide frequencies of that fragment. These frequencies are represented as a 16 dimensional vector:

$$[\ 0.056\ 0.071\ 0.067\ 0.05\ 0.081\ 0.058\ 0.056\ 0.057\ 0.051\ 0.051\ 0.069\ 0.07\ 0.049\ 0.079\ ]$$

summing up the first 8 components of this vector, we obtain 0.4999 which is the GC content of this fragment.

This basically is a linear projection on a line in the oligonucleotide frequency space which can be represented as the dot product of a vector with 4^k entries of 0s and 1s with an oligomer frequency vector. We can represent this mapping with P_{GC}, and the mapping operation as $f(X_{GC}) = P_{GC}(f(X))$, where $f(X_{GC})$ is the GC content, X is the k-mer relative frequency vector and $P_{GC}(.)$ is the linear function.

Summing up a trimer DNA composition vector with the help of the standard genetic code, we can approximately obtain the amino acid content vectors with a linear projection represented by a 20 X 64 binary matrix. Although the relative frequency of an amino acid equals the codon frequencies coding it, we substitute the codon frequencies with trinucleotide frequencies in order to obtain the relationship. The codon frequencies are calculated with a moving window of three bases, while the trinucleotide frequencies do not take the reading frames into account and average the frequencies over all reading frames. That is why this is an approximate mapping. The representation of this mapping is $(P_{X_{aa}} \circ P_{k3})$ and the mapping operation is $f(Xaa) \approx P_{aa}(P_{k3}(X_k))$, where $f(Xaa)$ is the amino acid content, $f(X)$ is the k-mer relative frequency vector, P_{aa} is the linear function mapping trimers to amino acid frequencies and P_{k3} is the linear function mapping k-mer frequencies to trimer frequencies. The error resulting from the approximation is negligible

$(r^2 = 0.9987, P < 0.0001)$.

Clearly, synoymous codon usage is obtained by the normalization of absolute codon frequencies (which are approximately the trinucleotide vectors) with the amino acid content. Both are linear projections in the oligonucleotide content space, which results in a nonlinear mapping within this space. The representation of this mapping is $(P_{scu} \circ P_{k3})$ and the mapping operation is $f(X_{scu}) \approx P_{scu}(P_k 3(f(X)))$, where X_{scu} is the vector containing synonymous codon usage, X is the k-mer relative frequency vector, P_{scu} is the nonlinear function mapping trimer frequencies to synonymous codon usage vectors and P_{k3} is the linear function mapping k-mer frequencies to trimer frequencies. There is a strong correlation between the approximate mapping and the actual synonymous codon usage values $(r^2 = 0.98, P < 0.0001)$.

Chapter 3

Other Computational Characterizations as Genome Signatures

The early genome signatures discussed in the previous chapter were defined by dinucleotide abundance ratios and Chaos Game Representations. Even though these two signatures were developed with different motivations and backgrounds, they share a significant common ground. Both classes of signatures can be defined as functions of oligonucleotide frequency vectors.

Here, we introduce other types of computational structures of DNA sequences, which can be categorized as genome signatures. As with the previously mentioned signatures, the computational characterizations which will be described here exhibit species specificity and pervasiveness. First present the mathematical characterization we wish to use as a signature. Then we test their specificity and pervasiveness. In practice, an ideal and absolute quality measurement to quantify the species specificity and pervasiveness of genome signatures is not currently known [65]. Nevertheless, it is possible to conduct relative comparisons based on the variation of certain parameters. For example, as the fragment size decreases, the computed genome signature will diverge from the signature derived from the entire genome. This deviation might vary based on the pervasiveness of a genome signature. Another example involving the relative species specificity of genome signatures is based on the similarity of genome sequences. Genomes of evolutionarily close organisms might be indistinguishable for some genome signatures, and they might turn out to be distinguishable using other signatures. This distinguishing ability is determined by the species specificity of a signature. There is no benchmark for specificity and pervasiveness against which to validate a mathematical structure as a genome signature. However, comparing the signatures based on these abilities, it is possible to have relative quantifications of pervasiveness and specificity. These can be obtained using statistical tests with varying genome fragment lengths at different taxonomy levels. We have used one way ANOVA statistics to measure the ratio of variance of signatures between the taxonomic levels to their variance within the taxa. This calculation is performed by F-measure. This constitutes our methodology to compare different mathematical characterizations of DNA sequences. Since all of the corresponding structures exhibit significant statistics (i.e., high F-values) implying pervasiveness and specificity, we refer to them as genomic signatures.

3.1 Long Term Correlation Statistics as Genome Signatures

Oligonucleotide frequency vectors consist of 4^k (k being the length of the oligomer) components, each component being the the frequency of a specific k-mer. Since the number of frequency parameters grows exponentially with the length of oligonucleotide, using long oligonucleotides results in data overfitting for average DNA fragment lengths. Therefore, oligonucleotide vectors of sufficient size are capable of capturing the short term dependencies in genomes. Thus genome signatures which are variants of oligonucleotide content (e.g., dinucleotide abundance ratios, chaos game representations) possess their signature characteristics due to the dependencies between nearby nucleotides. Long term correlations in DNA, on the other hand, also might be specific to the genome as well as being homogenous. If they are, and we could measure long term correlations in a genome, we could obtain computational genomic signatures. Observing long term correlations in DNA sequences may not be guaranteed, since it is not possible to find an intuitive rationale to propose conserved long term correlations in genomes. However, attempts to capture long term base dependencies can be made.

The correlation of a time series or a random process when the elements of the series are real numbers can be easily computed. For a wide sense stationary process the autocorrelation can be estimated as:

$$\hat{r}(k) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} (x_t - \mu)(x_{t+k} - \mu). \qquad (3.1)$$

Here, $x_t, n\mu, \sigma$ are the numeric sequence, its length, mean and variance respectively.

If biological sequences consisted of real numbers the profiles of $\hat{R}(k)$ could be tested for their signature characteristics. However, estimating the correlations of symbolic sequences is not that straightforward and requires either mapping to numerical sequences, or using models to represent the genomic sequences as symbolic random processes. The former approach has been used for representing DNA sequences as random walks. In a random walk model for DNA in which the walk is incremented by +1 if the next symbol is a pyrimidine (C, T) and decremented by -1 if it is a purine (A, G) base, the mean square fluctuation was observed to be different from that of a walk using random sequences or Markov models [66]. This is an indication of the existence of long term correlations in DNA. The existence of this correlation has been validated in various studies [67-71].

Investigating the correlations in genomes by mapping the DNA sequences into numerical data could provide an approach to study these dependencies. However, the results are dependent on the mapping and there is no trivial way of defining a mapping from a DNA sequences to a sequence of numbers. A more satisfactory approach is to use stochastic sequence analysis using the native alphabet. This can be done using concepts from information theory [72-75]. We first introduce an approach proposed by Dehnert et al. to estimate the long term correlations of DNA to be utilized as genomic signatures.

3.1.1 DNA as an Autoregressive Process

In a discrete autoregressive stochastic process, a symbol being emitted at time t is a function of the previous symbols. Therefore, the process has memory which can result in short-range, mid-range or long-range correlations. In most systems longer range correlations die out and become negligible in practice. Therefore the memory, or the order of the systems can be limited based on practical concerns. Autoregressive processes can be defined in terms of symbolic sequences. Such a model is called a discrete autoregressive process ($DAR(p)$) [76,77]. For a DNA sequence, where x_n is the n^{th} symbol, ($x_n \in \{A, C, G, T\}$) a $DAR(p)$ process can be defined as [78]:

$$x_n = V_n x_{n-A_n} + (1 - V_n) y_n. \tag{3.2}$$

Here V_n is a Bernoulli process taking values 1 with probability ρ and 0 with probability $1 - \rho$. A_n is an integer in $\{1, 2, 3, \ldots, p\}$, attaining each value with the probability $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_p$. y_n is another random process over the alphabet $\{A, C, G, T\}$ with independent and identically distributed probabilities for each n, represented by the marginal distribution π.

The process can be interpreted as follows. A new symbol in a DNA sequence is either picked from one of the previous p symbols, or selected independently. The process V_n works as a switch between random generation and selecting a symbol from near history. This solely depends on the random variable ρ. If ρ is zero, there are no dependencies between the nucleotides and DNA is a random sequence. At the other extreme, the sequence always depends on its context of length p. When the new symbol is picked from the previous p symbols, the probability α_i determines which symbol is to be selected. Note that α_i is the conditional probability of x_n being equal to x_{n-i} given x_n is selected from the history. Therefore, it can be used to model the dependencies of bases i positions apart in the sequence. That means the parameter vector $\alpha = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_p]$ can be used as a genome signature reflecting the dependencies of dinucleotides up to p bases apart.

Given the parameters of the $DAR(p)$ model, a simulated DNA sequence can be generated. However, to utilize this computational tool to define a genome signature, we have to estimate the parameters $\{\alpha_i\}$ given a DNA sequence. Dehnert et al. use a version of Yule-Walker estimation [78] to obtain the required parameters. According to this the autocorrelation function of the $DAR(p)$ process can be represented with the Yule-Walker equations:

$$r(k) = \rho\alpha_1 r(k-1) + \rho\alpha_2 r(k-2) + \ldots + \rho\alpha_p r(k-p), k \geq 1. \tag{3.3}$$

Expressing this as a system of linear equations we obtain:

$$\begin{aligned}
r(1) &= \rho\alpha_1 r(0) + \rho\alpha_2 r(1) + \ldots + \rho\alpha_p r(p-1) \\
r(2) &= \rho\alpha_1 r(1) + \rho\alpha_2 r(0) + \ldots + \rho\alpha_p r(p-2) \\
&\vdots \\
r(p) &= \rho\alpha_1 r(p-1) + \rho\alpha_2 r(p-2) + \ldots + \rho\alpha_p r(0)
\end{aligned}$$

Given the autocorrelation values, this set of equations can be solved and $\alpha = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_p]$ can be obtained. It has been shown that the ad-hoc autocorrelation estimator performs well with symbolic sequences [77]. In this case the autocorrelation function is:

$$\hat{r}(k) = 1 - \sum_{a_i \in S} B_m(k, a_i) \frac{1}{1 - \pi(a_i)}. \tag{3.4}$$

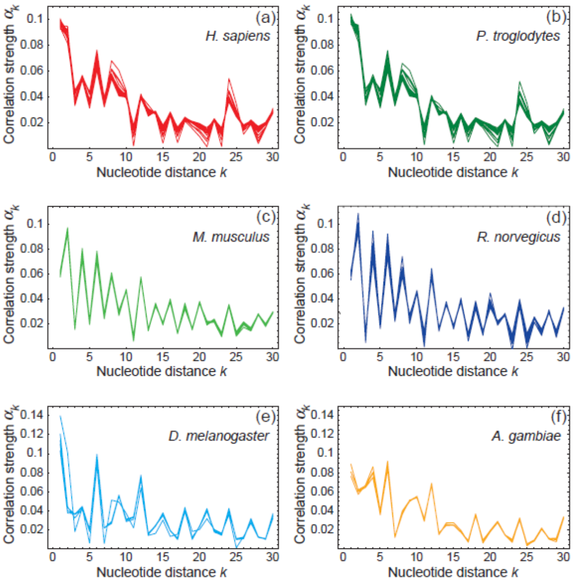

Figure 3.1: Correlation strength profiles for the first 30 components for *H. sapiens*, *P. proglodytes*, *M. musculus*, *R. norvegicus*, *D. melanogaster*, and *A. gamblae*.

Here $S = \{A, C, G, T\}$, and the function B_m is

$$B_m(k, a_i) = \frac{1}{m-k} \sum_{a_i \neq a_j \in S} \sum_{l=1}^{m-k} \delta_{a_i}(x_l)\delta_{a_j}(x_{l+k}) \qquad (3.5)$$

where $\delta_a(x) = 1$ when $a = x$ and 0 else.

With this version of Yule-Walker estimation of $DAR(p)$ model parameters, the estimated vector $\alpha = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_p]$ can be used as a genome signature. This particular signature has been used for modeling eukaryote chromosomes and measuring distances between chromosomes of the same organism and chromosomes from different organisms [79,80]. In Figure 3.1 the plots of α vectors of dimension thirty are illustrated for all chromosomes of 6 eukaryotic organisms.

It is visually evident that while the intergenomic parameter vectors are very similar, the pattern is different for different organisms. This implies the species specificity and pervasiveness of the α vectors [79], and thus it is a genomic signature. This genomic signature has been reported to be quite specific, but it becomes hard to distinguish the signatures between closely related species. Using an ℓ_1 metric (i.e. $d(\alpha_1, \alpha_2) = \sum_i |\alpha_1(i) - \alpha_2(i)|$) to measure the distance of signatures, it was observed that the chromosomes of human and chimpanzee are difficult to distinguish from each other.

3.1.2 Average Mutual Information Profiles

Another method of detecting long range correlations in DNA sequences is the use of average mutual information. Average mutual information was first introduced by Claude Shannon for the study of signals under noisy channel conditions [81]. It has attracted the attention of computational biologists as a means for understanding dependent events like correlated mutations at noncontiguos sites [82], and secondary structures and correlations in protein sequences [83-90].

Assume x is a random process emitting the DNA sequence, where x_i and x_j are instances corresponding to the bases in DNA. The information about x_i contained in x_j and vice versa is given by:

$$\begin{aligned} I(x_i; x_j) &= H(x_i) - H(x_i|x_j) \\ &= H(x_i) - (H(x_i, x_j) - H(x_j)) \qquad (3.6) \end{aligned}$$

where $H(.)$ is the Shannon entropy. In this case, a DNA sequence is again viewed as a stochastic process with the assumption that the process is wide sense stationary and ergodic. Those two assumptions imply that information about base distributions can be estimated from DNA fragments and they are not position dependent. That is to say, the dependency of base pairs at a fixed distance apart does not depend on the positions of the individual bases but just on the distance between the bases. Therefore, the entropies can be estimated over x and an average information can be assigned for nucleotide pair placed k bases apart for all $|j-i|=k$. Then the average mutual information can be written as:

$$\begin{aligned}
I(x_i;x_j) &= I(x;x(k)) = I(k) && (3.7)\\
&= H(x) + H(x(k)) - H(x,x(k)) &&\\
&= -\sum_i P(x_i)\log_2(P(x_i)) - \sum_i P(x_{i+k})\log_2(P(x_{i+k})) && (3.8)\\
&\quad +\sum_i P(x_i,x_{i+k})\log_2(P(x_i,x_{i+k})) &&\\
&= \sum_i P(x_i,x_{i+k})\log_2(\frac{P(x_i,x_{i+k})}{P(x_i)P(x_{i+k})}). && (3.9)
\end{aligned}$$

The probability estimations can be simply done by relative frequency counts of the pairs located k base pairs apart. The estimate of the average mutual information gives a statistical measure of how much information is shared between nucleotides k bases apart. Therefore, it forms a measure of the correlation within a DNA sequence. When the profile of a set of location distance values such as $[I(1)I(2)\ldots I(n)]$ is compiled, that forms an average mutual information profile (AMI profile), providing the dependencies in short-, mid- or long-range. AMI profiles appear to be different for varying species and the signatures obtained from different parts of organisms resemble each other. Bauer et al. [88] investigated the

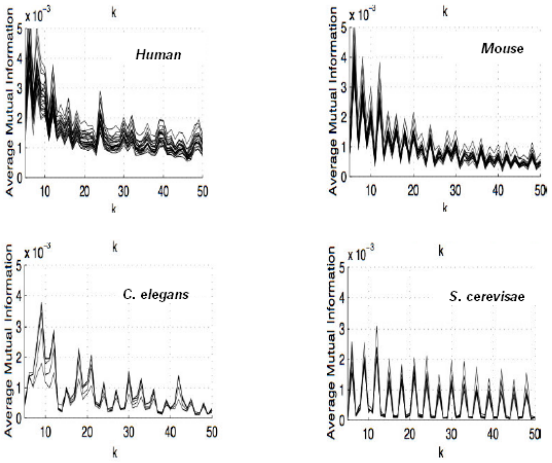

Figure 3.2: Average Mutual Information profiles for the first 50 components for *H. sapiens*, *M. musculus*, *C. elegans*, and *S. Cerevisae.*(figure taken from Bauer et al [88])

signature behavior of AMI profiles and observed that just like correlation strength, AMI profiles are similar for different chromosomes of the same eukaryotic organism. They also showed that AMI profiles show different patterns for each of those organisms. These two properties imply the species specificity and pervasiveness of AMI profiles. In Figure 3.2, AMI profiles to $n = 50$ are plotted for all the chromosomes of four eukaryotic organisms, the species specificity and pervasiveness can be observed graphically from the figure.

We have argued that most genomic signatures belong to the same class, because they can be deduced from long oligonucleotide counts, and they are all functions of oligonucleotide occurrence. Even though there is such a relationship between the correlation strength signature and the oligonucleotide occurence, it is not explicit

Fvalue	1 kbp	10 kbp	50 kbp
Genus	105.75	708.65	2215.9
Family	120.08	818.15	2544.4
Order	127.28	875.54	2664.7

Table 3.1: F-scores of one-way-ANOVA for AMI profiles. Distribution of profiles for varying fragment length at different clade levels are considered.

since the estimation of correlations are performed via the parameter estimation of a discrete autoregressive model. On the other hand, this relation can still be claimed for AMI profiles. We can view $I(k)$ as the average log odds ratio of dinucleotide relative abundance, where the dinucleotides are located k bases apart from each other. It can be shown that the corresponding frequencies can be obtained by linearly projecting oligonucleotide count vectors. Assume the oligonucleotide frequency vector for n-mers:

$$f(x_1x_2\dots x_n) = \begin{bmatrix} f(AAA\dots A) \\ f(AAA\dots C) \\ \vdots \\ f(TTT\dots T) \end{bmatrix}. \tag{3.10}$$

the AMI profile is calculated over the dinucleotide frequencies. It is possible to deduce the dinucleotide frequencies from the genome signature by aggregating entries by summing them up. The resulting vector with 16 entries is:

$$f(x,x(k)) = \begin{bmatrix} f(xx\dots xAx\dots xAx\dots x) \\ f(xx\dots xAx\dots xCx\dots x) \\ \vdots \\ f(xx\dots xTx\dots xTx\dots x) \end{bmatrix}. \tag{3.11}$$

Here x denotes a wildcard variable which represents any of the four bases. Clearly $(k-1)$ $f(x, x(k))$ vectors can be generated from the oligonucleotide count vectors by summing over the wildcard variables x. This operation corresponds to a matrix multiplication of the n-mer frequency vector with a $n \times 4^n$ vector of 1s and 0s. Therefore the AMI profile can be considered to be a nonlinear function of a linear projection in the oligonucleotide frequency space which consequently is a nonlinear mapping of the oligonucleotide count vectors. However, these theoretical results do not have an important implication in practice. This is because the required dimension for that mapping requires a very high number of parameters that cannot be estimated with realistic genome sizes. For example, the number of parameters for an oligonucleotide vector to deduce an AMI profile of 30 variables, is around 360 million fold greater than the total length of the human genome. The corresponding estimation would result in overfitting with relative frequency counts. Therefore, both measures can be assumed as belonging to a different class of genome signatures that utilize the longer range correlations in genomes.

The F-values of AMI profiles for ANOVA tests are provided in Table 3.1. Fragment lengths of 1 kbp, 10 kbp, and 50 kbp are used for the clade levels of genus, family, and order. High F-values indicate that AMI profiles can be considered as genome signatures.

The factors resulting in the conservation of longer range correlations in DNA are not well understood. Long range dependencies are mostly attributed to the structural properties of DNA such as supercoiling and the corresponding 10-11 bp periodicities [89,90]. Also Alu and SINE repeats [91] and tandem repeats are thought to result in long range correlations. However, removing all annotated repeats and investigating the correlation strength signature, it is still possible to observe the intragenomic similarities [92]. This behavior might be an imply that structures other than well-known repeats are involved in the long-range correlation process.

3.2 Signatures Based on Composition Vectors

Clearly, it is possible to define many different types of composition vectors using different functions of oligonucleotide content. Moreover, several of them exhibit significant species specificity while being sufficiently conserved within a genome. We will briefly review a subset of them which make sense in terms of representing over- and underabundance of oligonucleotide usage or representing the short term dependencies in DNA.

3.2.1 Markov Models

Markov models have been used frequently in order to detect intragenomic heterogeneities. Primarily, models trained on coding and noncoding sequences were employed to predict gene sequences from open reading frames [93]. Different evolutionary pressures create compositional differences in genes and intergenic regions on an intragenomic scale and Markov models are able to distinguish between the compositional differences of coding and noncoding regions. Intuitively, we expect Markov models to capture global compositional features in intergenomic scale. This was noted by Salzberg and colleagues [94] who used a variable-order Markov model based gene prediction program for the classification of genomic sequences from different organisms.

As genomic signatures,we can view the Markov models as being the conditional probability of a base given its finite length context. Here DNA sequences are assumed to be stationary random processes, and the probability of a base is independent of the bases located outside the context of that base, i.e. the process has finite memory. Thus, the conditional probability can be written as:

$$p(x_i|x_{i-1}x_{i-2}\ldots) = p(x_i|L_i) \tag{3.12}$$

where L_i is the context of the base x_i. This context can be of different lengths for different bases, which result in variable order Markov models. Fixing the length of L_i generates fixed order Markov models. In the case of fixed order Markov models, the model parameter can be estimated as the ratio of two different sized oligomer counts:

$$\begin{aligned} p(x_i|x_{i-1}x_{i-2}\ldots) &= p(x_i|x_{i-1}x_{i-2}\ldots x_{i-k}) \\ &= \frac{p(x_{i-k}x_{i-k+1}\ldots x_{i-1}x_i)}{p(x_{i-k}x_{i-k+1}\ldots x_{i-1})}. \end{aligned} \tag{3.13}$$

For a k^{th} order Markov model every base has 4^k different context. The profile of 4^{k+1} different parameters can form a genomic signature. For the same context, the probabilities of the four bases sum up to one, therefore, the last one can be calculated from the other three. The genome signature profile, thus, has 3×4^k free parameters.

Dalevi et al. used Markov models as genomic signatures and showed that these signatures are more specific than oligonucleotide counts [95]. By estimating variable order Markov models they repeated their experiments and reported a slight

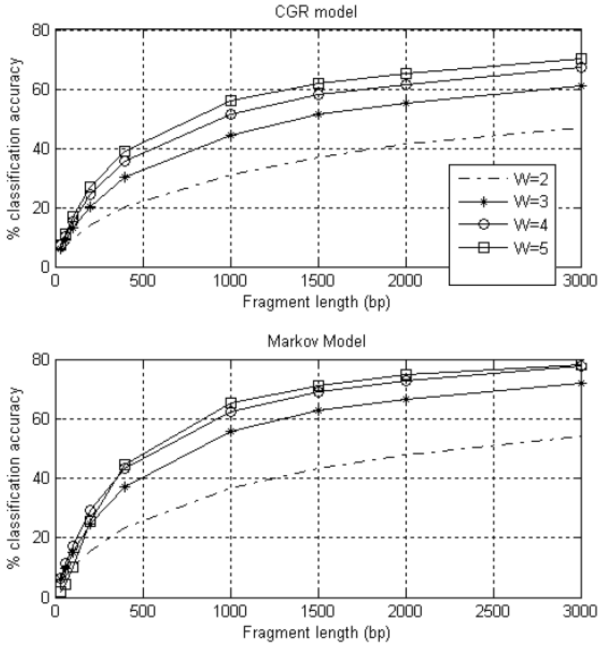

Figure 3.3: Comparison of true positive ratios for CGR signatures and Markov models based on multiple hypothesis testing. Random genome fragments shorter than 3000 bp are used. The tests are repeated for oligonucleotide length from dimers to pentamers.

improvement in species specificity with this modification.

In Figure 3.3, the multiple hypothesis testing true positive ratios are plotted for different short genomic fragment lengths. Comparing the Markov models with oligonucleotide frequencies of the same order, it can be seen that Markov models are more species specific for all oligonucleotide lengths.

The F-values of Markov models for ANOVA tests are provided in Table 3.2. Fragment lengths of 1 kbp, 10 kbp, and 50 kbp are used for the clade levels of genus, family, and order.

Fvalue	1 kbp	10 kbp	50 kbp
Genus	312.88	2891.7	9079.99
Family	328.91	3041.5	9307
Order	364.82	3402.8	10004.2

Table 3.2: F-scores of one-way-ANOVA for Markov model parameters. Distribution of profiles for varying fragment length at different clade levels are considered.

3.2.2 Abundance Profiles of Oligonucleotides

The heavy-tailed behavior of k-distributions implies a significant over- and under-abundance of oligomers within a genome. This is an indication of dependencies of nearby nucleotides, which results in a deviation of their frequencies of occurrence from the expected values. We can expand Karlin's abundance measurement scheme based on Markov assumption to general k-mers.

Consider a k-mer $x_1, x_2, \ldots, x_k$. with probability $p(x_1, x_2, \ldots, x_k)$. We can write this probability as:

$$p(x_1, x_2, \ldots, x_k) = p(x_k|x_1, x_2, \ldots, x_{k-1})p(x_1, x_2, \ldots, x_{k-1}) \tag{3.14}$$

We can rewrite the first factor on the right hand side of Equation (4.9) under different independence assumptions as follows. Assuming that the bases occur independently of each other the conditional probability can be replaced by the marginal probability

$$p(x_k|x_1, x_2, \ldots, x_{k-1}) = p(x_k) \tag{3.15}$$

Now we can calculate the odds ratio of an oligonucleotide frequency, and its expected value based on this zeroth order Markov model

$$cv_0(x_1, x_2, \ldots, x_k) = \frac{p(x_1, x_2, \ldots, x_k)}{p(x_k)p(x_1, x_2, \ldots, x_{k-1})} \tag{3.16}$$

If we assume that the bases follow a first order Markov model

$$\begin{aligned} p(x_k|x_1, x_2, \ldots, x_{k-1}) &= p(x_k|x_{k-1}) && (3.17) \\ &= \frac{p(x_{k-1}, x_k)}{p(x_{k-1})} && (3.18) \end{aligned}$$

The corresponding relative abundance index rai_1 is then given by

$$cv_1(x_1, x_2, \ldots, x_k) = \frac{p(x_1, x_2, \ldots, x_k)p(x_{k-1})}{p(x_{k-1}, x_k)p(x_1, x_2, \ldots, x_{k-1})} \tag{3.19}$$

If the particular k-mer occurs more frequently than would be predicted based on the first order Markov model $rai_1(x_1, x_2, \ldots, x_k)$ will be greater than one, otherwise it will be less than one; the magnitude depending on how far the actual distribution of the oligomer varies from the prediction of the model. Continuing in this fashion we obtain

$$\begin{aligned} cv_2(x_1, x_2, \ldots, x_k) &= \frac{p(x_1, x_2, \ldots, x_k)p(x_{k-2}, x_{k-1})}{p(x_{k-2}, x_{k-1}, x_k)p(x_1, x_2, \ldots, x_{k-1})} && (3.20) \\ cv_3(x_1, x_2, \ldots, x_k) &= \frac{p(x_1, x_2, \ldots, x_k)p(x_{k-3}, x_{k-2}, x_{k-1})}{p(x_{k-3}, x_{k-2}, x_{k-1}, x_k)p(x_1, x_2, \ldots, x_{k-1})} && (3.21) \\ \vdots \qquad & \vdots \\ rai_{k-2}(x_1, x_2, \ldots, x_k) &= \frac{p(x_1, \ldots x_k)p(x_2 \ldots x_{k-1})}{p(x_2, \ldots x_k)p(x_1, x_2, \ldots x_{k-1})}. && (3.22) \end{aligned}$$

Therefore, a general scheme for calculating the deviation of oligonucleotide frequencies based on Markov models of order i ($i < (k-1)$) can be defined using cv_m: the ratio of joint distribution of k-mers over m^{th} order Markov expansion. The F-values of compositional vectors of lowest and highest orders for ANOVA tests are provided in Table 3.4 and Table 3.5, respectively. Fragment lengths of 1 kbp, 10 kbp, and 50 kbp are used for the clade levels of genus, family, and order.

3.2.3 Abundance Profiles Based on Zero'th Order Markov Model Frequency Estimations

The oligonucleotide abundance profiles defined previously are based on estimating the frequencies of oligonucleotides using Markov assumption. Although the models vary in the order of Markov models adopted, they all assume dependence of adjacent bases in a sequence. Zero'th order Markov model estimation differs in the calculation of expected frequencies. According to this abundance calculation, the frequency of an oligonucleotide is determined by the frequency of each base. The bases are independent and identically distributed within the genome. Thus, no correlations exist within a sequence. The abundance value is calculated as follows:

$$ZOM(x_1, x_2, \ldots, x_k) = \frac{p(x_1, x_2, \ldots, x_k)}{p(x_1)p(x_2)\ldots p(x_k)}. \tag{3.23}$$

This profile is a measurement to determine how much each oligonucleotide diverges from random distribution. ZOM's are known to carry strong phylogenetical signals [189], from which consistent phylogenetic trees can be constructed. It was reported that this taxonomic classification ability is comparable with 16s RNA phylotyping, indicating significant species specificity. In fact, dinucleotide abundance ratio profiles is in this class. ZOM calculations for tetranucleotide frequencies have been

Fvalue	1 kbp	10 kbp	50 kbp
Genus	244.81	2515	7192.9
Family	254.81	2574.4	7903.4
Order	255.5	2713	7569.3

Table 3.3: F-scores of one-way-ANOVA for Zeroth order Markov model profiles. Distribution of profiles for varying fragment length at different clade levels are considered.

Fvalue	1 kbp	10 kbp	50 kbp
Genus	135.59	1911	5274.4
Family	143.88	2042.3	6208.6
Order	160.67	2326.4	6513.9

Table 3.4: F-scores of one-way-ANOVA for cv_0 profiles. Distribution of profiles for varying fragment length at different clade levels are considered.

frequently used as a genome signature and it has been accepted to be a successful genome signature [197]. The one way ANOVA tests in Table 3.3 indicates that zero'th order Markov models constitute a strong genome signature class.

3.3 Oligonucleotide Frequency Derived Error Gradient (OFDEG)

The signatures described to this point are either related to the short-term or the medium term dependencies of DNA sequences and they are expressed as profiles.

Fvalue	1 kbp	10 kbp	50 kbp
Genus	48.6	819.53	2327.5
Family	54.7	932.37	2517.4
Order	64.69	1102.9	2977.8

Table 3.5: F-scores of one-way-ANOVA for cv_{k-2} profiles. Distribution of profiles for varying fragment length at different clade levels are considered.

These profiles are elements of a multidimensional space. The oligonucleotide frequency derived error gradient (OFDEG), on the other hand, is a scalar genome signature calculated based on the convergence rate of oligonucleotide frequencies estimation with increasing sequence length [96]. The biological foundation of this signature has not been explored. However, in practice OFDEG is observed to be very species specific and pervasive although it is represented with only a single parameter.

The oligonucleotide frequencies in a genomic fragment are clearly a better estimate of the oligonucleotide content than the estimate gathered from a subsequence of this fragment. Due to ergodicity assumption, as the number of samples (i.e., longer fragment length) increases, the estimations converge asymptotically. OFDEG simply attempts to capture this convergence behavior by subsampling the fragment and measuring the decrease in error as the length of the subsamples increases up to the fragment length.

The derivation of OFDEG is as follows: for a given fragment, the oligonucleotide frequencies of length k is calculated and stored at the OF_{full} vector. Starting with an initial subsequence length L_1, random p subsequences are drawn from the given fragment and the oligonucleotide frequency is calculated over each subfragment. The errors in the frequency counts are stored as

$$e_{1,j} = OF_{full} - OF_{L_1,j}, \tag{3.24}$$

where $j \in \{1, 2, \ldots, p\}$. Increasing the subfragment length by l bp, p subsequences

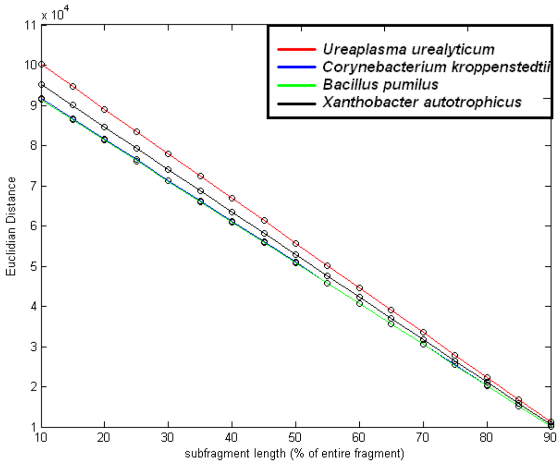

Figure 3.4: The errors of frequency counts are plotted *U. urealyticum*, *C. kroppenstedtil*, *B. pumilus*, and *X autoptropicus* are plotted. The linear decays imply that each organism attains a specific gradient (i.e. OFDEG) value.

are sampled at each iteration and the errors are calculated in the same fashion:

$$e_{i,j} = OF_{full} - OF_{L_1+il,j}, \tag{3.25}$$

where $i \in \{1, 2, \ldots, n\}$ and n is the last iteration number determined by the subfragment length reaching some percentage of the original fragment (typically 80%). The relation of increasing subfragment length and corresponding decreasing error is observed to be a linear decay. The last step of the OFDEG calculation is the measurement of the gradient of this decay using linear regression. The slope of the regression line gives the characteristics of the genome and is used as a genome signature. In Figure 3.4 the relationship is plotted for different genomes where the species specificity of the decay gradient can be observed. For a comprehensive set

Fvalue	1 kbp	10 kbp	50 kbp
Genus	133.34	679.88	1740.5
Family	137.11	699.01	2034.1
Order	161.76	825.29	2442.9

Table 3.6: F-scores of one-way-ANOVA for OFDEG values. Distribution of profiles for varying fragment length at different clade levels are considered.

of prokaryotes, using multiple hypothesis testing by classifications, it was observed that the true positive detection ratios of OFDEG derived from tetranucleotide frequencies are comparable to the specificity of tetranucleotide frequency vectors for the genomic fragments around the range of 8 bkp [96].

The ANOVA of some of the signatures introduced in this chapter are performed using similar tests performed for amino acid usage and synonymous codon usage in chapter 2. Different fragment lengths varying between 1 kbp and 50 kbp are used in the tests. The results can be seen in tables 3.1-3.6. The major observation of ANOVA tests is that simpler models result in clearer separation in the Euclidean space they are placed.

Chapter 4

Measuring distance of biological sequences using genome signatures

4.1 Introduction

We have viewed computational genomic signatures as mathematical structures mapped from DNA sequences to a metric space. Throughout the discussion of computational genomic signatures, we have focused on the two basic signature features; specificity and pervasiveness. The former determines the distinguishability of different genomes, the latter determines its usefulness when only fragmentary information about the genome is available. In order to develop efficient applications of the genome signature concept in a number of computational biology problems, strongly species specific and pervasive characterizations are required. In this chapter, we introduce a methodology to efficiently exploit the information gathered by genome signatures efficiently.

Different mathematical characterizations of DNA fragments emphasize different specific features native to each genome. It is possible to observe that based on the previously classified mathematical characterizations. Oligonucleotide frequency counts estimate the occurrence probability of each oligomer in a sequence. Markov models quantify the emission probability of each nucleotide, based on the short-term context of the corresponding base. Similarly, abundance profiles measure the divergence of a sequence from randomness, which is related to the complexity and organization at genome level. Oligonucleotide frequency derived error gradient is another measure that quantifies the genome complexity by investigating the frequency count change with varying fragment length. Average mutual information and correlation strength signatures, on the other hand, characterize a genome using longer term correlations in DNA sequences.

All the features measured by the corresponding genome signature representations are characteristic to each genome, and thus they are species specific. However, their species specificity might be different, and they might exhibit different pervasive natures. An absolute quality measurement for genome signatures is hard to define. However, using fundamental statistical tests on the signatures sampled from existing genomes helps us to compare signatures and determine their relative power. We have observed that different characterizations of DNA leading to different signatures vary in relative quality. This can be attributed to the capability of the signature to emphasize specific signals and capture native structural properties, as well as the homogeneity of the captured features. The one-way ANOVA tests performed on different genome signatures gives idea about the relative quality of the signatures. The calculated F-values measure how distinguishable the distributions of signatures from different sources are. Since the variances in F-value calculations are derived from the sum of squared distances, the genome signature space can be considered as a Euclidean space. We have empirically observed that simple characterizations have greater quality than the signatures attempting to capture genome structure in a more sophisticated way. For example, oligonucleotide count vectors attain higher F-values than Markov model parameter vectors. Moreover, Markov models attain higher F-values than oligonucleotide abundance profiles which are estimated using Markov models.

Regarding the empirical distributions of different genome signatures in Euclidean space, an appropriate strategy appears to be employing simple models such as oligonucleotide frequency vectors. This has been a main strategy in various computational biology applications. However, the structure of signatures is not the only factor determining their quality. How we interpret the signatures quantitatively also effects their specificity and pervasiveness. Different distance measures can affect the utility and, therefore, the power of a signature. It may be possible to better differentiate genomes with the same signature depending on how we measure distances between signatures. In this sense, the mathematical characterization of DNA sequences and the metrics proposed to compare these characterizations are both components of genome signatures. We have denoted the mathematical characterization as the signatures; because with distances other than standard norms, the signature becomes implicit. The defined metric maps the characterizations obtained from the Euclidian space to another metric space and calculates the distance in the mapped space. However, since this mapping is not necessarily explicit, we cannot have a representation of the corresponding signature; and thus we stick with the definition in two parts, *genomic signature +*

distance measurement, as the total characterization.

To exemplify the importance of the distance measure, consider the oligonucleotide frequency signatures with two different distance metrics. A thousand random genomic fragments from 99 prokaryotic genomes (each one picked from a different genus) were sampled for different fragment lengths. The signature used to represent each fragment was the vector of pentanucleotide frequencies. It is expected that signatures for fragments from the same genomes are similar to each other, and signatures of fragments from different genomes are different from each other. Therefore, we expect the signatures of fragments from the same genome to be clustered together in the composition space. As a result of this clustering, it is possible to classify these signatures using supervised classification algorithms. We chose maximum margin classifiers, performed ten-fold cross-validation and measured the ratio of true positives, which is a measure for the quality of the signature. The results were obtained by repeating the test with two similarity measures. First, we measured the similarity (S_1) of two signatures as the dot product of 5-mer vectors, where f_i and f_j represent the pentamer frequency vectors for genome fragments i and j. Second, the similarity S_2 was obtained via the Gaussian Kernel. The two distance metrics are:

$$\begin{aligned} S_1(f_i, f_j) &= f_i^T f_j \qquad (4.1) \\ S_2(f_i, f_j) &= exp(-\frac{\parallel f_i - f_j \parallel}{\sigma}) \end{aligned}$$

It can be shown that using the Gaussian kernel in this manner is equivalent to mapping the input vector into an infinite dimensional space and taking the inner

Similarity metric	400 bp	1000 bp	2000 bp	5000 bp
S_1	51.3%	71.1%	75.3%	82.9%
S_2	69.5%	81.1%	87.5%	91.8%

Table 4.1: The ten-fold cross validation results of 1000 genomic fragments gathered from 99 genera for varying fragment length. The % true positive ratios are supplied.

product in that space [192]. That is:

$$S_2(f_i, f_j) = exp(-\frac{\parallel f_i - f_j \parallel}{\sigma}) = \Phi_\infty(f_i)^T \Phi_\infty(f_j). \tag{4.2}$$

The results with the two similarity measures are shown in Table 4.1. There is a significant difference in the accuracy of classification, and the species specificity and pervasiveness obtained using the Gaussian kernel is clearly superior. We can view this result in two different ways. Because of the kernel duality, the Gaussian similarity metric can be assumed to be the inner product of the signatures, $\Phi_\infty(f_i)$ and $\Phi_\infty(f_j)$. The superiority is because the implicit signatures, $\Phi_\infty(f_i)$ and $\Phi_\infty(f_j)$, capture the characteristics of the sequence better; or the improvement can be attributed to the similarity measurement and because the kernel similarity exploits the signals in the same genome signature better than the dot product calculation. In either case, the importance of the distance measure is evident.

Before the introduction of the RAIphy method, we will review a number of difference/similarity measurement schemes in the context of particular signatures. The total operation can be interpreted as implicit signatures with better characterization of genomes; or it can be interpreted as the use of measures exploiting the information embedded in the same signature better than the conventional measures.

4.2 Classical Methods: Euclidian Distances and Correlation Statistics

The most popular sequence similarity/distance measurement for genome signatures are based on the simple ℓ-norms or correlations of signature profiles expressed as vectors in a multidimensional space. A well known example of this is the δ-distance measure of Karlin et al., which is a version of the ℓ_1 norm:

$$\ell_1(S_1, S_2) = \sum_i |S_1(i) - S_2(i)|, \tag{4.3}$$

Karlin et al. used the δ-distance measure with genomic signatures based on the di-, tri- and tetranucleotide abundance vectors [47-55]. The ℓ_1 norm was also utilized in the calculation of short-range and midrange correlation strength profiles [78-80]. In order to quantify the differences and similarities between the CGRs of different sequences Deschevanne et al. [62] defined the Euclidian distances of images calculated from pixel differences, which actually corresponds to the Euclidian distance of oligonucleotide frequency vectors [62-64]

$$\ell_2(S_1, S_2) = \sum_i (S_1(i) - S_2(i))^2. \tag{4.4}$$

A machine learning methodology based on unsupervised neural networks called self-organizing maps has been employed for the purpose of clustering short genomic fragments of the same origin together with the help of genomic signatures [180-183]. Self-organizing maps use Euclidian distance in the training of neurons, thus these methods can be considered to be in the class of genomic signatures used with Euclidian distances. OFDEG signatures [96] have also been used with the same metric.

Another popular technique to measure DNA sequence similarity is the Pearson correlation of genomic signature profiles:

$$\rho_{S_1,S_2} = \frac{\sum_i (S_1(i) - \overline{S_1(i)})(S_2(i) - \overline{S_2(i)})}{\sqrt{\sum_i (S_1 - \overline{S_1(i)})^2}\sqrt{\sum_i (S_2 - \overline{S_2(i)})^2}}. \tag{4.5}$$

The Pearson coefficient has been used to calculate the similarities of abundance profiles of k-mers calculated over $(k-1)^{th}$ order Markov model expectations [193,194], as well as of abundance profiles of k-mers calculated over zeroth order Markov model expectations [195-197].

Correlation measurements and ℓ-norms perform well with the general characteristics of genome signatures. However, they obscure pervasive signals by averaging out genome-wide total signals. This phenomenon can be observed better in short genomic fragments. Consider a genome with certain oligomers that are either over- or underrepresented. This characteristic is expected to be homogeneously represented within the genome. Yet, gathering statistics from a short genomic fragment is perhaps not sufficient to compile a full profile of word preferences. Assume a fragment of length 200 bp where the genome signature is determined to be 7-mer frequencies. The profile sampled from this genomic fragment will only represent 200 7-mers and they may be observed several times (i.e. <200 words will be observed). The total number of possible 7-mers are $4^7 = 16,384$; and in this case only around 1% of the full profile is represented. Nevertheless, the 7-mers with nonzero frequency of occurrence are mostly from the set of overrepresented 7-mers, which means there are detectable pervasive signals even with an insufficient number of samples. However, comparing all possible words in the distance/similarity calculation gives weight to the nonrepresented, and because there are many more nonrepresented oligomers; this makes detecting the pervasive signals due to the overrepresented oligomers more difficult. The classical comparison methods of ℓ-norms and correlation coefficients are in the class of metrics taking all 4^7 signature parameters into account. Clearly, an adaptive comparison metric that takes only represented words into account could have done better.

4.3 Distances Based on Model Fitness

We have briefly discussed an inherent weakness of the classical genome signature similarity/distance measurement approach. Now, we introduce some relative similarity measurement approaches based on model fitness which are potentially better at exploiting pervasive genome signature signals and at mitigating the problems that occur with the classical distance metrics. This is particularly true in applications where it is necessary to detect the genome of origin of short genomic fragments. In most applications, the use of absolute distance/similarity metrics limits the employment of relatively longer oligonucleotide counts. Since all 4^k k-mer frequencies are involved in those similarity measurements, good estimates of all of these k-mer frequencies is necessary. This requirement implies that overfitting should be avoided since overfitting would dramatically drop the detection accuracy. In order to prevent overfitting, more data points are needed in order to accurately estimate frequencies of occurrence (i.e., longer fragments); and the number of parameters to be estimated should preferably be kept small (i.e., shorter oligonucleotides of length k where 4^k is equal to the number of different oligonucleotides). Not surprisingly, all these methods work best with 4-mers and 5-mers with sequences $\geq$ 1-3 Kbp. However, longer-range correlations exist in DNA sequences which we would like to exploit for characterizations even with short sequence reads.

4.3.1 Likelihood Functions

The first class of similarity measures we discuss can be viewed as likelihood functions of oligonucleotide probabilities estimated from relative frequency counts. This class of similarity measures was first used by Sandberg et al. [169] for detecting the species of origin for short genomic fragments of unknown source. In this setting, given a genomic fragment the probability of a genome being the origin of the corresponding fragment is calculated as $P(G_i|f)$ where G_i is the i^{th} genome in a set of organisms and f is the genomic fragment. The genome resulting in the highest probability ($\arg max_i P(G_i|f)$) is determined to be the origin of this fragment. According to Bayes' Theorem:

$$P(G_i|f) = \frac{P(f|G_i)P(G_i)}{P(f)}. \tag{4.6}$$

If the prior probability of observing a genome is assumed to be equal for all organisms, the source genome is determined to be the genome G_i which would result in the highest probability of observing the fragment. In terms of genome signatures, this is simply the probability of emitting the genome fragment f, with the oligonucleotide probabilities estimated from genome i. Assuming independence of different oligonucleotides, this probability calculation turns out to be the multiplication of related oligomer probabilities:

$$P(f, G_i) = \prod_x P_{G_i}(x_1, x_2, \ldots, x_k)^{n_f(x_1, x_2, \ldots, x_k)} \tag{4.7}$$

where $P_{G_i}(x_1, x_2, \ldots, x_k)$ is the relative frequency of occurrence of the oligonucleotide $x_1 x_2 \ldots x_k$ computed from genome G_i, and $n_f(x_1, x_2, \ldots, x_k)$ is the number of times that oligonucleotide occurs in the fragment f. This probability estimate provides a measure of the likelihood that a fragment has been obtained from a particular genome based on the oligonucleotide content. Note that only the oligonucleotides observed in fragment f are involved in the calculation (i.e., nonobserved oligomers do not contribute to the product). This implies that the nonobserved oligomers are filtered resulting in the capture of pervasive signals and elimination of the noise stemming from the use of statistics of words not in the sample. Using this relative distance, Sandberg et al. [169] were able to substantially reduce the size of the fragments that could be accurately classified obtaining a 90 percent classification accuracy for fragments of size 1.5 kbp in a set of 28 prokaryotic genomes from various genera.

In Figure 4.1, the comparison of this measure with Pearson correlation and Euclidian distance is shown for 7-mer frequency signatures with various genomic fragment lengths. The usage of a relative measure increases the quality of the signature in total being more specific for all short fragment lengths. Dalevi et al. [170] extended the work of Sandberg et al. by replacing the probabilities in Equation 4.7 with conditional probabilities and variable-order Markov models. This turns out to be:

$$P(f, G_i) = \prod_x P_{G_i}(x_k | L_k)^{n_f(x_k | L_k)} \tag{4.8}$$

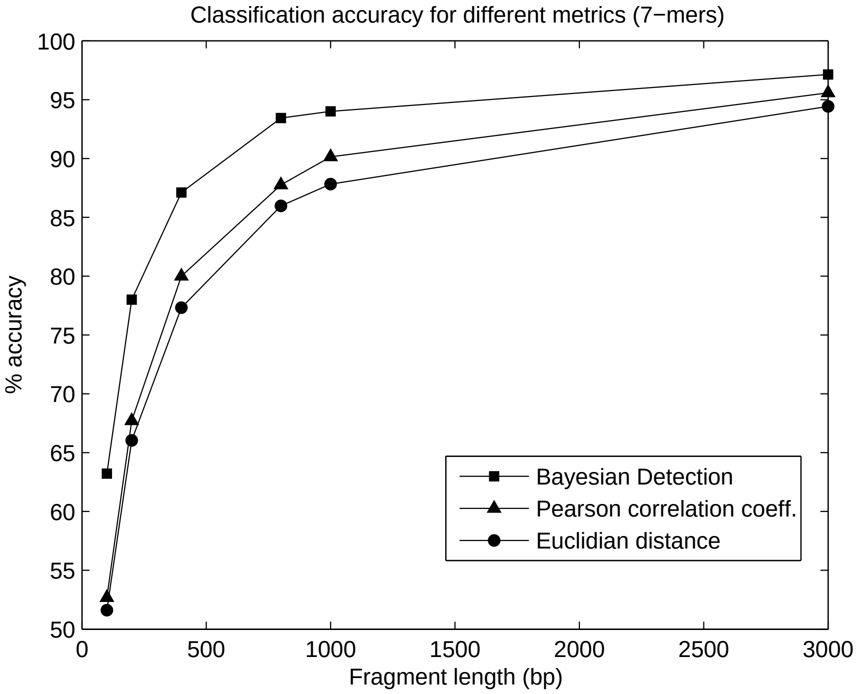

Figure 4.1: The accuracy performance of different distance/similarity metrics for 28 taxa with varying fragment length is shown. The frequencies of 7-mers are used. Using the metric defined by Sandberg et. al. appears to be more accurate for all fragment lengths than employing Euclidian distance and Pearson correlation coefficients.

where L_k is the context of the k^{th} base x_k determined by the Markov model trained on the genome G_i. L_k is the (k-1)-mer $x_1x_2\dots x_{k-1}$ if the Markov model is of fixed order. Improvement over the likelihood function calculated by oligonucleotide content was reported [170], which is an improvement in the quality of the signature resulting from the change in profile (i.e., employing conditional probabilities instead of oligomer probabilities in the signature).

4.3.2 Indexing Based on Oligonucleotide Abundance

The same idea of using relative measures which consider only observed words in a short genomic fragment can be extended to other signatures. In turn, signatures emphasizing the over- and underabundance of oligonucleotides can be modeled in a profile and used as an index. Subsequently, the average scores attained by the oligonucleotides observed in a short genomic fragment can be used as a similarity measure. The abundance calculation for a k-mer can be obtained using an l^th order Markov assumption ($l < k$). In this section we describe such an indexing scheme which we call the relative abundance index (RAI).

In order to build a comprehensive abundance index it is useful if a combination of different order Markov models contribute to the characterization. We accomplish this in the following manner. First, we use models of various orders to predict the frequency of occurrence of the k-mer under consideration. We then use the log of the ratio of the observed frequency to the predicted frequency to provide an indication of how well or how poorly the k-mer follows the various Markov models.

Consider a k-mer $x_1, x_2, \dots, x_k$ with probability $p(x_1, x_2, \dots, x_k)$. We can write this probability as:

$$p(x_1, x_2, \ldots, x_k) = p(x_k|x_1, x_2, \ldots, x_{k-1})p(x_1, x_2, \ldots, x_{k-1}) \tag{4.9}$$

We can rewrite the first factor on the right-hand side of Equation (4.9) under different independence assumptions as follows. Assuming that the bases occur independently of each other the conditional probability can be replaced by the marginal probability:

$$p(x_k|x_1, x_2, \ldots, x_{k-1}) = p(x_k) \tag{4.10}$$

To test this assumption, we can compute a log-odd ratios as in Karlin et al. to form the RAI of order 0 rai_0:

$$rai_0(x_1, x_2, \ldots, x_k) = \log_2 \frac{p(x_1, x_2, \ldots, x_k)}{p(x_k)p(x_1, x_2, \ldots, x_{k-1})} \tag{4.11}$$

If we assume that the bases follow a first order Markov model,

$$\begin{aligned} p(x_k|x_1, x_2, \ldots, x_{k-1}) &= p(x_k|x_{k-1}) && (4.12) \\ &= \frac{p(x_{k-1}, x_k)}{p(x_{k-1})} && (4.13) \end{aligned}$$

The corresponding relative abundance index rai_1 is then given by:

$$rai_1(x_1, x_2, \ldots, x_k) = \log_2 \frac{p(x_1, x_2, \ldots, x_k)p(x_{k-1})}{p(x_{k-1}, x_k)p(x_1, x_2, \ldots, x_{k-1})} \tag{4.14}$$

If the particular k-mer occurs more frequently than would be predicted based on the first-order Markov model, $rai_1(x_1, x_2, \ldots, x_k)$ will be positive, otherwise it will be negative. The magnitude will depend on how far the actual distribution of the oligomer varies from the prediction of the model. Continuing in this fashion we obtain:

$$\begin{aligned}
rai_2(x_1, x_2, \ldots, x_k) &= \log_2 \frac{p(x_1, x_2, \ldots, x_k)p(x_{k-2}, x_{k-1})}{p(x_{k-2}, x_{k-1}, x_k)p(x_1, x_2, \ldots, x_{k-1})} && (4.15)\\
rai_3(x_1, x_2, \ldots, x_k) &= \log_2 \frac{p(x_1, x_2, \ldots, x_k)p(x_{k-3}, x_{k-2}, x_{k-1})}{p(x_{k-3}, x_{k-2}, x_{k-1}, x_k)p(x_1, x_2, \ldots, x_{k-1})} && (4.16)\\
&\vdots \quad \vdots\\
rai_{k-2}(x_1, x_2, \ldots, x_k) &= \log_2 \frac{p(x_1, \ldots x_k)p(x_2 \ldots x_{k-1})}{p(x_2, \ldots x_k)p(x_1, x_2, \ldots x_{k-1})} && (4.17)
\end{aligned}$$

We can combine the RAIs of all orders by adding them to give:

$$rai(x_1, x_2, \ldots, x_k) = \sum_{i=0}^{k-2} rai_i(x_1, x_2, \ldots, x_k) \tag{4.18}$$

Given a particular k-mer $x_1, \ldots, x_k$, $\{rai(x_1, x_2, \ldots, x_k)\}$ gives an indication of how well the k-mer follows a Markov model. The smaller the model is that can predict the frequency of occurrence of the k-mer, the smaller will be the value of $\{rai(x_1, x_2, \ldots, x_k)\}$. For example, if the $k-$mer followed a third-order model but not a lower order model, one would expect the RAIs of an order greater than or equal to three to have a value close to zero. If the k-mer can only be explained by a fifth order model and not by a model of order less than 5, then one would expect more of the coefficients to deviate from zero. In particular, k-mers that occur “unexpectedly” would have a high relative abundance index for all models and thus a high value in the sum of Equation (4.20). In this manner $\{rai(x_1, x_2, \ldots, x_k\}$ identifies oligomers that vary significantly from a set of Markov models.

4.3.3 The Specificity of RAI Characterization

Lemma: DNA fragments belonging to the same generalized source with a given RAI profile are expected to have higher RAI scores than the DNA fragments of another source.

Proof: This observation is fundamental to our similarity measure. To observe this situation, we assume that K-mer frequencies from a group follows the same probability distribution for the DNA sequences in this group and different groups follow other probability distributions.

If the group α follows the K-mer probability distribution P_α, then the expected RAI score for the fragments of this group for the RAI profile of the same group turns out to be

$$E_F[rai^{G_\alpha}] = \sum_{\mathbf{x}} P_\alpha(x_1, x_2, \ldots, x_k) rai^{G_\alpha}(x_1, x_2, \ldots, x_k), \tag{4.19}$$

recalling that the RAI profile $rai^G(x_1, x_2, \ldots, x_k)$, derived from the training sequence of the group α, is

$$rai^{G_\alpha}(x_1, x_2, \ldots, x_k) = \sum_{i=0}^{k-2} rai_i^{G_\alpha}(x_1, x_2, \ldots, x_k). \tag{4.20}$$

Therefore, the RAI score is

$$E_F[rai^{G_\alpha}] = \sum_{\mathbf{x}} P_\alpha(x_1, x_2, \ldots, x_k) \sum_{i=0}^{k-2} rai_i^{G_\alpha}(x_1, x_2, \ldots, x_k), \tag{4.21}$$

Plugging in the RAI profile definition for the α fragment,

$$\begin{aligned}
E_{F_\alpha}[rai^{G_\alpha}] &= \sum_{i=0}^{k-2}\sum_{\mathbf{x}} P_\alpha(x_1,\ldots,x_k)\log_2 \frac{P_\alpha(x_1,\ldots x_k)P_\alpha(x_{k-1}\ldots x_{k-i})}{P_\alpha(x_k,\ldots x_{k-i})P_\alpha(x_1,\ldots x_{k-1})} \\
&= \sum_{i=0}^{k-2}\{\sum_{\mathbf{x}} P_\alpha(x_1,x_2,\ldots,x_k)\log_2 P_\alpha(x_1,x_2,\ldots x_k) \qquad (4.22)\\
&+ \sum_{\mathbf{x}} P_\alpha(x_1,x_2,\ldots,x_k)\log_2 P_\alpha(x_{k-1}\ldots x_{k-i}) \\
&- \sum_{\mathbf{x}} P_\alpha(x_1,x_2,\ldots,x_k)\log_2 P_\alpha(x_k,x_{k-1},\ldots x_{k-i}) \\
&- \sum_{\mathbf{x}} P_\alpha(x_1,x_2,\ldots,x_k)\log_2 P_\alpha(x_1,x_2,\ldots x_{k-1})\}.
\end{aligned}$$

We can simplify the equation (4.22) using the entropy definition $H(p) = -\sum P\log P$ and the marginalization property that $\sum_{x,y} P(x,y)\log P(y) = \sum_y \log P(y)\sum_x P(x,y) = \sum_y P(y)\log P(y)$. Then

$$\begin{aligned}
E_{F_\alpha}[rai^{G_\alpha}] &= \sum_{i=0}^{k-2}\{\sum_{\mathbf{x}} P_\alpha(x_1,x_2,\ldots,x_k)\log_2 P_\alpha(x_1,x_2,\ldots x_k) \qquad (4.23)\\
&+ \sum_{\mathbf{x}} P_\alpha(x_{k-1}\ldots x_{k-i})\log_2 P_\alpha(x_{k-1}\ldots x_{k-i}) \\
&- \sum_{\mathbf{x}} P_\alpha(x_k,x_{k-1},\ldots x_{k-i})\log_2 P_\alpha(x_k,x_{k-1},\ldots x_{k-i}) \\
&- \sum_{\mathbf{x}} P_\alpha(x_1,x_2,\ldots x_{k-1})\log_2 P_\alpha(x_1,x_2,\ldots x_{k-1})\} \\
&= -\sum_{i=0}^{k-2}(H_k(P_\alpha) - H_{k-1}(P_\alpha) + H_i(P_\alpha) - H_{i-1}(P_\alpha))
\end{aligned}$$

where $H_i(P_\alpha)$ stands for the i^{th} order entropy. Summing up the telescopic summation we obtain

$$\begin{aligned} E_{F_\alpha}[rai^{G_\alpha}] &= -\sum_{i=0}^{k-2}(H_k(P_\alpha) - H_{k-1}(P_\alpha) + H_i(P_\alpha) - H_{i-1}(P_\alpha)) \qquad (4.24) \\ &= -(k-1)H_k(P_\alpha) - (k-2)H_{k-1}(P_\alpha) - H_{k-1}(P_\alpha) + H_{k-2}(P_\alpha) \\ &= -(k-1)H_k(P_\alpha) - (k-2)H_{k-1}(P_\alpha) - h_{k-1}(P_\alpha). \end{aligned}$$

In the equation (4.24), $h_{k-1}(P_\alpha)$ stands for the conditional entropy where $(k-2)$ previous bases are given in a (k-1)-mer.

Following similar steps, the RAI score obtained for a β fragment is evaluated as

$$\begin{aligned} E_{F_\beta}[rai^{G_\alpha}] &= \sum_{i=0}^{k-2}\{\sum_{\mathbf{x}} P_\beta(x_1, x_2, \ldots, x_k)\log_2 P_\alpha(x_1, x_2, \ldots x_k) \qquad (4.25) \\ &+ \sum_{\mathbf{x}} P_\beta(x_{k-1}\ldots x_{k-i})\log_2 P_\alpha(x_{k-1}\ldots x_{k-i}) \\ &- \sum_{\mathbf{x}} P_\beta(x_k, x_{k-1}, \ldots x_{k-i})\log_2 P_\alpha(x_k, x_{k-1}, \ldots x_{k-i}) \\ &- \sum_{\mathbf{x}} P_\beta(x_1, x_2, \ldots x_{k-1})\log_2 P_\alpha(x_1, x_2, \ldots x_{k-1})\} \end{aligned}$$

for the each entropy component in the equation (4.24), we can use the property that average self information of a distribution is smaller than the average information obtained by a different distribution: [198]

$$H(P_\alpha) = -\sum P_\alpha \log(P_\alpha)) \leq -\sum P_\beta \log(P_\alpha)) \qquad (4.26)$$

Thus, each component of the equation (4.24) is greater than the corresponding components of the equation (9). We obtain

$$E_{F_\alpha}[rai^{G_\alpha}] \geq E_{F_\beta}[rai^{G_\alpha}] \tag{4.27}$$

■.

Empirical Distributions of Membership Scores

Here, histograms of the Relative Abundance Index scores are shown for different levels of phylogenetic closeness. A RAI profile is built for a species and RAI scores calculated using this profile for a relatively close relative and a distant relative is considered. A close relative is expected to have higher RAI scores and a lower score is expected for a distant relative. The histograms are derived over 10000 random samples of 400 bp DNA fragments.

We observed the RAI scores with fragments from varying phylogenetical relations. Figure 4.2 shows RAI score distributions of DNA sequences from relatively close sources. The first set of fragments belong to another strain of a species from which the RAI profile is calculated. The second set of fragments are from another species in the same genus. The score distributions are observed to be close. Figure 4.3 shows RAI score distributions of DNA sequences from moderately distant sources. The first set of fragments belong to another species of a genus from which the RAI profile is calculated. The second set of fragments are from another genus in the same family. The score distributions are observed to be differing moderately. Figure 4.4 shows RAI score distributions of DNA sequences from distant sources. The first set of fragments belong to another species of a genus where the RAI profile is calculated from. The second set of fragments are from another phylum. The score distributions are observed to be differing significantly.

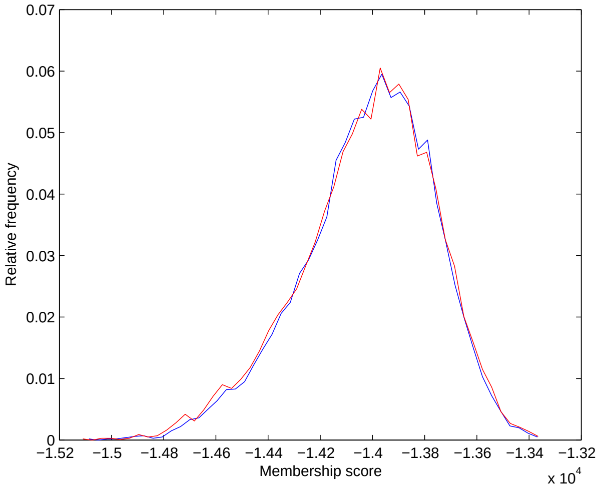

Figure 4.2: RAI profile is derived from *Salmonella enterica subsp. enterica serovar Typhi Ty2*. Blue histogram: Scores of *Salmonella enterica subsp. enterica serovar Typhi str. CT18* fragments, red histogram: Scores of *Salmonella typhimurium*fragments. Species from same genus show very close behaviors with the RAI profile in the same genus.

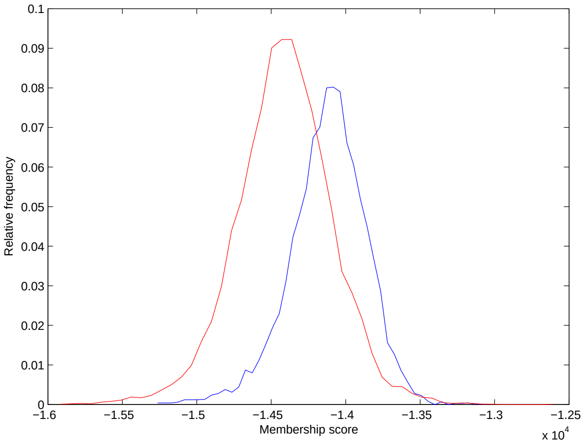

Figure 4.3: RAI profile is derived from *Chloroflexus sp. Y-400-fl.* Blue histogram: Scores of *SChloroflexus aggregans* fragments, red histogram: Scores of *Roseiflexus sp. RS-1*fragments. All species are from Chloroflexaceae family. The RAI profile and first set of fragments are from Chloroflexus genus where the second set of fragments belong to another genus, Roseiflexus. Fragments from moderately distant relatives show a moderate difference in RAI scores.

We have seen that similarity/distance measurement between mathematical characterizations of DNA fragments is a factor determining the strength of a genome signature as well as the structure of the characterization. Therefore, a more appropriate distance measurement could result in better distinguishability of DNA fragments. This observation also provides the opportunity of using more complicated mathematical characterizations as genome signatures. Although, these models have potential to capture more information from genome sequences, absolute distance metrics fail to exploit this information. We have developed a genome signature (RAI) which combines the divergence of oligonucleotide frequencies from their expected values, estimated by different orders of Markov assumptions. The similarity of relative abundance values are measured using a probabilistic framework. The new signature is capable of modeling a genome better than currently

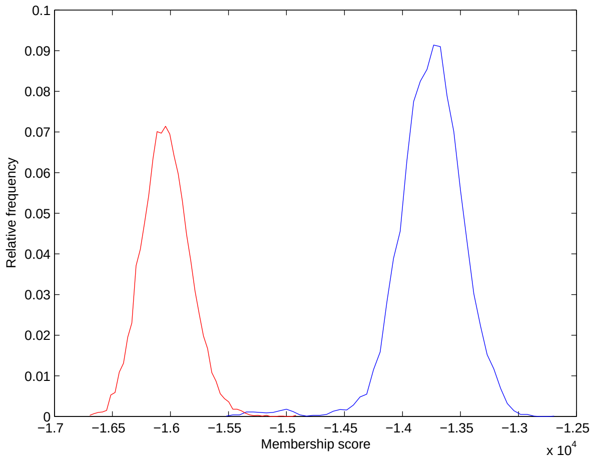

Figure 4.4: RAI profile is derived from *Staphylococcus aureus*. Blue histogram: Scores of *Staphylococcus saprophyticus* fragments, red histogram: Scores of *Pseudomonas aeruginosa* fragments. The RAI profile and first set of fragments are from Staphylococcus genus of Firmicutes, where the second set of fragments belong to another phylum, Proteobacteria. Fragments from distant relatives show a significant difference in RAI scores.

known genome signatures. As a result, RAI attains better assignment accuracy of unknown genome fragments to their origin of species. The strength of RAI as a signature makes it a powerful candidate as an approach to a metagenomics problem called taxonomic binning. The background for the field of metagenomics and related problems are reviewed in the following chapter. Subsequently, applications of RAI signature for metagenome binning will be introduced.

Chapter 5

Metagenomics Background

5.1 Community Analysis of Environmental Samples

Microbic organisms are involved in numerous processes of life on Earth. Microorganisms are a source of nutrients, cycling organic matter, and they form symbiotic relationships with life forms at every level of the tree of life. In aggregate they make up a great proportion of the living population in the biosphere. While the microbial world dominates life on Earth and understanding of this world is crucial for many areas ranging from biological sciences to other fields such as medicine, agriculture or food production, our current understanding of microbes is very limited. It is estimated that less than one per cent of the microbial world has been explored [97,98]. This is primarily due to the technical limitations on isolation and culturing of microbes in nature. Only a small percentage of microbes can be cultured and studied by microbiologists. Thus, the current knowledge of microbiology is biased in favor of the small proportion of culturable species.

Since the sequencing of the first bacterial genome in 1995 [99], genomes of more than 1000 microbial species have been sequenced and annotated. This number is much less than the known minority of microbial diversity. Since it is not possible to isolate the majority of existing microbes, the current paradigm is not sufficient for extensively exploring the tree of life. Naturally it brings the problem of limiting genomic analysis to the small percentage of the existing species which are culturable. The newborn science of **metagenomics**, often acknowledged as a paradigm shift in microbiology [100], has the potential to overcome the limitations on microorganism annotation. Metagenomics enables the genomic study of environmental samples: and thus, it deals with the unknown majority of microbes for which isolation of single genomes is not possible [101, 102]. A principal goal of metagenomics is the sampling of microbiomes and recovery of the genetic material without the isolation of single organisms.

Recovering the genetic material *en masse* provides great opportunities for various areas of research. *In situ* sampling enables recovery of genetical material from various environments such as ocean [103,104], soil [105], hot springs and hydrothermal vents [106], polar ice caps [107], and hypersaline environments [108]. This new type of complex data gathered from the environment directly requires novel analysis approaches as it introduces new research challenges. However, even the early techniques involving conventional genome analysis has revealed valuable insights. Exploring the taxonomic and metabolitic diversity at the ecosystem level is one of the practical achievements of metagenomics. Analysis of environmental samples also leads into advances in biotechnology [109,110], the study of human physiology [111], and genetical archeology of extinct species [112,113]. Discovery of novel genes for encoding biocatalysts and drugs, as well as the discovery of other biomolecules can be counted as the achievements of the early era of metagenomics [114-116]. Eventually, advances in metagenomics should help to extend the tree of life [117] while enriching sequence libraries. Furthermore, the study would expand analysis from genomic to metagenomic: interactions within communities could be studied extensively using samples from various habitats.

5.2 Sampling and Sequencing Environmental Samples

In order to gather genetic material from an environmental sample, the first step is to sample organisms from the environment. The goal of metagenome sampling is to obtain sufficient number of chromosomes from each species existing in the microbial community. Population sizes of different operational taxonomic units (OTUs) in a mixture might be diverse, resulting in the underrepresentation of low populated species. This imposes a requirement on the amount of chromosomes that should be gathered from the environment in order to achieve a complete representation of the community.

Rarefaction curves, which plots the number of OTU's gathered versus the number of individuals sampled, are used to determine the quality of sampling [118]. As the slope of a rarefaction curve converges to zero, a complete representation of the microbial population is obtained. Ideally, many individuals can be sampled to guarantee a complete sampling of the metagenome. However, due to the increased cost and restricted budgets of metagenomics projects, an ideal sampling might not always be possible.

Following sampling, environmental samples are filtered. This is a physical procedure, and organisms in a sample are eliminated according to their physical size. The goal in many microbiology projects is to eliminate small viroids and large protists to obtain the sampled bacterial population in the corresponding habitat. There are other metagenome projects that are targeted to viromes [119], and in this case viral organisms are subject to filtering process.

Whole shotgun sequencing of the recovered organisms is the next step required to obtain the genetic information. The product of sequencing depends on factors such as the sampling size, and the sequencing technology employed.

Depending on the diversity of the microbial community, an environmental sample can include from a few dominant species to thousands of species at the same level of dominance [120-122]. Examples of low diversity metagenomes include the gutless worm symbiont community [123], for which long contigs in the range of 100 kbp - 1 Mbp were assembled, and acid mine drainage biofilms [120], in which complete genome assemblies of the dominant species were obtained. However, in diverse communities only very short contigs are achievable. In termite hindgut microbiomes [124], and soil and whale fall (deep ocean) [125] samples, contig assemblies do not exceed 10 kbp in length.

This missing data problem stems mostly from the sequencing constraints and project budgets. For the popular Sanger sequencing method [126, 127] metagenomic projects usually result in a total of 100 Mbp [128]. For a community of microbes with different abundance ratios, this amount of data will only cover relatively abundant sequences while the rest of the population will remain with insufficient coverage roughly proportional to their relative abundance in the population.

A worse scenario exists for high diversity communities: none of the organisms will have enough coverage for the assembly of long contigs. This results in missing portions of the genomes in the sample and short sequences which are generally insufficient for analysis of genes and phylogenetic diversity [129, 130].

The Lander-Waterman equation [131] suggests that generation of longer total sequenced data will proportionally increase the average coverage per base. Given a properly sampled environmental sample, this would mean sufficient coverage to assemble organisms with lower abundance is possible in theory with production of massive amounts of sequencing output. With the introduction of high-throughput sequencing technologies, lower cost per base and faster sequencing is now possible [132-134]. Second generation sequencing technologies are replacing high-cost and labor-intensive Sanger sequencing. The Life Sciences 454-GS FLX Titanium 454 pyrosequencer [135] can produce 400 Mbp in a single run while the Illumina GAIIx [136] can produce 15-20 Gbp per run, the $SOLiD^{TM}$ (Sequencing by Oligo Ligation and Detection) platform [137] can yield 20 Gbp per run and the single-molecule sequencing platform, $HelicosHeliScope^{TM}tSMS$ [138] is capable of producing >1Gbp/hour. The feasibility of producing greater amounts of metagenome data has accelerated the area of metagenomics. It was reported that in the last 5 years, second generation sequencing has generated a greater amount of sequenced DNA than Sanger sequencing has generated in the last three decades [129].

5.3 Exploration of Biodiversity in a Metagenome

For ideal phylogenetic and functional genomics analysis, complete genomes are needed. In practice, fragmented genomes in long contigs can also be very informative for various levels of analysis. However, this is only currently achievable for dominant species in low diversity populations. This lack of ability to obtain sufficiently long contigs from individual genomes in a microbial mixture has forced researchers to approach the metagenome as a "bag of genes" and conduct the analysis on a gene level. Phylogenetic diversity is usually explored by characterizing OTUs using polymerase chain reaction (PCR) amplification of marker genes such as 16S rRNA genes [139] or using non-rRNA genes [130, 141].

Multiple housekeeping genes are used in Multilocus Sequence Typing (MLST) for exploring the phylogenetic diversity [142]. Unfortunately, approaches which estimate phylogenetic diversity using marker genes are known to have several problems [143, 144]. Recently a core set of marker genes were determined to be used in phylotyping. AMPHORA [145] and MLTreeMap [146] analyze these marker genes to infer the phylogenetic information of a given environmental sample. While these programs supply information about the biodiversity of a sample, they only associate those genome fragments that carry a marker gene with possible OTUs. This means that the great majority of sequencing reads remain unassociated with any taxa. Table 5.1 shows the percentage of the DNA sequences in a metagenome mixture which are assigned to taxa using phylotyping methods. The reason for this poor assignment is that only a small part of a genome contains the marker genes, and this dramatically reduces the occurrence probability of a marker gene for a given random genome fragment. In fact phylotyping approaches suggest answers for the question "what groups are in the mix?" rather than the question of "Which fragment belongs to which one of those groups in the mixture?"

Method	16s RNA	MLST	AMPHORA	MLTreeMap
Assignment (%)	< 1	< 1	1.3	1.89

Table 5.1: The percentage of fragments assigned to taxa in a metagenome using marker gene-based phylotyping methods.

5.4 Metagenome Assembly

Metagenome assembly is the process of obtaining long contigs or drafts of complete genomes from sequence reads. The sequenced metagenomes include fragment reads of multiple genomes from various organisms existing in the environment. An ideal scenario for the assembly of genomes populating the metagenome would be assembling each genome in parallel fashion after a taxonomical classification phase [147, 148]. Realizing such an approach is currently an open research problem.

The contemporary approach to metagenome analysis is to employ taxonomic grouping after attempts to assemble the metagenome treating it as a single species read set. There are several problems with this approach. Taxonomic classification operates successfully with the sequences having a length in the long contig range [58, 59]. On the other hand, attempting to assemble an entire metagenome without taxonomic grouping, or binning, leads to poor assemblies. This is the conundrum of metagenomics data analysis: a good assembly of genomes in an environmental sample requires phylogenetic classification, while good phylogenetic classification requires assembled contigs of sufficient length and thus, containing significant information for characterization. To date no comprehensive metagenome assembler has been reported and conventional genome assemblers are facing difficulties with data consisting of a mixture of several genomes which eventually affects the performance of taxonomic classification. Currently, single genome assembly programs such as Forge, Phrap [149], TIGR, CAP3 [150], Arachne [152, 152], JAZZ [153], the Celera Assembler [154], and EULER [155, 156] are also employed for metagenome assembly [157]. These programs are specifically designed for Sanger sequencing and the assembly of isolated genomes. Modifications to these algorithms adapting them to perform on the greater number of shorter reads yielded by new generation sequencing are also available with the programs such as SSAKE [158], VCAKE [159], SHARCGS [160], Velvet [161] and Allpaths.

5.5 Metagenome Binning

Binning one of the computational tasks in metagenome analysis, involves categorizing sequenced data into operational taxonomic units (OTUs) for further analysis. Binning is a difficult problem when the information required for differentiation has to be obtained from short DNA reads. A number of approaches has been proposed for computational binning of metagenome data, and some of them are currently employed in real-life metagenome analysis.

It is possible to categorize the binning approaches in three main classes: similarity search methods, supervised compositional methods and unsupervised methods. While the first category involves molecular database searches for previously explored homogenous sequences, the latter two use the notion of genome signatures to bin the DNA sequences to taxa.

5.5.1 Similarity Search-Based Binning Methods

Probably the most widespread method of binning is using homology searches for a given unknown genomic fragment. As mentioned earlier, using a few marker genes is insufficient to label a great majority of metagenomical fragments. However, employing larger sets of molecular sequences is shown to serve the purpose of metagenome binning. Here, larger sets of molecular sequences refer to comprehensive sets of protein sequences and assemblies of whole genomes or large contigs from the known organisms. Corresponding molecular data gathered from various projects are deposited in public databases. Consequently, the task of searching for matches between unknown metagenome samples and known sequences reduces to homology searches in molecular databases.

An example of employing homology search using known protein domains is the algorithm Carma [162]. Carma assigns sequences to taxonomical origins by trying to match them to known protein families contained in Pfam domains. Profile Markov models are used to search the aligned Pfam domains for possible homologies. Although this class of methods is frequently used for phylotyping, they can be employed for binning since they comprehensively compare protein domains and attempt to classify any given genome fragment. While computationally expensive, Carma has been shown to be accurate even for short sequences in the current pyrosequencing read length range (80-400 bp). However, the accuracy drops dramatically when phylogenetically close sequences are missing from the search databases. Running CARMA on a comprehensive dataset gathered from a large spectrum of known genomes resulted in inaccurate classifications [162]. (6% sensitivity when using 100 bp sequences for identification at the genus level).

Another similarity based method is MEGAN [163, 164], which uses the scores of similarity searches to assign the DNA fragments to taxa using a lowest common ancestor algorithm. Usually nucleotide BLAST [165] is employed as the similarity search task. Therefore, a common binning strategy using MEGAN appears to be a local alignment search using available DNA sequences of known organisms. MEGAN is reported to be successful when the organisms forming the metagenome have close relatives in the search databases. However, in a recent study [166], only 12% of the data obtained from microbial communities in coral atolls got significant BLAST hits. SOrt-ITEMS [167] is a recent example of similarity search methods employing BLAST as ontology search strategy. In addition to similarity search scores, the search parameters are also considered in the taxonomy assignment algorithm.

Similarity search methods are very powerful when the homologous sequences exist in search databases, because significant hits with local alignments are expected to have high ratios of true positives. On the other hand, homology searches would be unable to identify sequences from a large proportion of the microbial population. The reason behind this incapability is the small ratio of sequenced biological molecules compared to the vast number of species in metagenome samples. As a matter of the course, poor identification results are reported with real-life metagenome data.

5.5.2 Supervised Compositional Binning Methods

Supervised compositional binning methods approach the problem of binning from a general perspective of modeling. According to this scheme, genome fragments are represented as compact mathematical models which represent the species specific characteristics of genomes. Sequenced genomes in public databases are also represented by their models. The homology search task of sequence similarity-based methods is replaced with model comparison. The model based approach provides several advantages: first, the computational burden is reduced when compared to the similarity based methods, and second the models provide a more general representation. The reduction in computational burden is a crucial practical issue in metagenomics analysis, since large amounts of data have to be processed, which might result in infeasibility problems. Similarity based methods require sequence alignment runs over voluminous databases. Whereas, supervised compositional binning methods generally compare relatively small structures. Moreover, the representation of sequences by structures that emphasize the specific features provides a concise framework. Introduction of a more general scheme has been observed to be more accurate for a number of binning scenarios [168].

Genome signatures, being species specific and pervasive, are a plausible candidate for DNA sequence modeling to be employed in supervised binning methods. While the specific character helps in distinguishing fragments from different genome sources, the pervasiveness enables the use of the signature with short fragments usually seen in metagenomes.

A naive-Bayesian Classifier-based method proposed by Sandberg et al. [169] and a Markov chain method by Dalevi et al. [170] are early examples of this approach. The algorithm PhyloPythia [171] consists of various support vector machine (SVM) classifiers. Relative frequency profiles of short oligonucleotides (5-mers for clade levels of genus to class, and 6-mers for the clade levels of phylum and domain) were used as feature vectors. Relative oligonucleotide frequency vectors were generated for various fragment lengths and SVMs were trained using different fragment lengths. Satisfactory sensitivity and specificity results are reported for the sequence lengths $>$ 1-3kbp. However, a sharp cut-off in the accuracy is observed for fragments less than 1 kbp in length. Another recent taxonomic classification method, TACOA [172], proposes a k-nearest neighbor classification based algorithm. In this method, genomic sequences are represented by over- underabundance profiles of oligonucleotides called genomic feature vectors (GFV). GFV's are identical to zero'th order Markov models. Training GFV's over known genomes, the best score calculated from the closest k trained neighbors to a test GFV determines the taxonomic assignment of an unknown test query. For sequence lengths under 1 kbp, 4-mers are used to build GFV's. For longer sequences, the frequencies of 5-mers are observed to perform the best. TACOA has been shown to correctly classify fragments larger than 800 bp with an average sensitivity between 76% at the rank of superkingdom and 39% at the rank of genus. Its performance is comparable to PhyloPythia in that range. As the distance metric, Euclidean distances are used and fed into radial basis functions in PhyloPythia, whereas inner products are used in TACOA.

Phymm [168] was developed for the classification of short read lengths of metagenomics data. It is based on a Bayesian decision machine which detects the taxonomic source of a read with its maximum a posteriori probability calculated over variable order Markov models. Complete genomes of known taxa are used for training Markov models. Oligonucleotide lengths of 1-mer to 8-mer are used in training the models. Phymm shows significantly increased accuracy compared to CARMA and PhyloPhytia.

5.5.3 Unsupervised Methods

The previous two classes of binning methods require prior knowledge of sequence information for known taxonomic units. When the majority of species embodying a metagenome is included in model or sequence databases, the binning performance is satisfactory. When unidentified and non-sequenced genomes exist in the mixture, the taxonomic classification becomes impossible. Given contemporary limited knowledge of microbial sequences, this is not an unexpected scenario. Furthermore, discovery of new microbes is conceptually very limited with the similarity-search based and supervised methods. Since supervised and similarity-based binning methods label the metagenome with known species, the exploration is confined to the small portion of the known microworld, or its close relatives.

For discovery of novel microbial species, unsupervised categorization of metagenomes is needed. The requirement for unsupervised binning is the ability to distinguish fragments of different sources without the aid of trained models. That is to say, accurate clustering of metagenome samples has to be achieved. Employing genome signatures within an autonomous framework of categorization appears to be an appropriate approach to unsupervised binning.

Unsupervised Binning Using Self Organizing Maps

Early examples of unsupervised binning made use of autonomous neural network structures called self organizing maps (SOM) [173,174]. SOM's group similar structures using batch learning methods which minimize the mean square classification error. SOM's are useful for the visualization of high-dimensional data; they project the complex relation of data onto a simple two dimensional map.

The possibility of clustering metagenome samples using genome signatures was extensively investigated in [175]. It was previously reported that genomes sharing the same environment are similar in composition [176-179]. As organisms in a metagenome share the same environment this could result in a problem of disappearance of species specific features of genome signatures in metagenomes. However, a case study performed on an acid mine metagenome in which the organisms share extremely acidic conditions has shown that the genome signatures are not obscured. Using SOM's as the clustering scheme and tetranucleotide frequencies of 5 kbp fragments a clear clustering of metagenome samples were observed. Moreover specificity was observed for fragments as short as 500 bp, and clusters form around the length of 1400 bp.

Abe et al. [180] reported a clear separation of species with 1 kb and 10 kb fragments from 65 prokaryotes and 6 eukaryotes using 2,3,4-mer oligonucleotide frequencies. They also supported their results using clinical data from uncultured microbes [180]. Comparing the clusters with the known genomes, they concluded that 79% of the Sargasso Sea metagenome consists of unknown species.

Different architectures of SOM's further improved the binning results of this class of unsupervised methods. Using growing self organizing maps, hyperbolic SOM's in unsupervised [181,182] and semi-supervised settings [183], accuracy values comparable with supervised binning were achieved.

Binning Methods Considering Community Abundance

The diversity of populations and under-overabundance of species in a microbial community affect the clustering characteristics of metagenome binning. If a taxon has an abundant number of individuals the variance of signatures within the taxon might be large, compared to inter-taxa variance of low abundance sequences. Different approaches which take into account the population abundance have been implemented in a number of binning programs.

Compostbin [184] uses data reduction with weighted principal component analysis. The 4^6 dimensional feature space of hexanucleotide frequencies calculated for each fragment is reduced down to three dimensions of largest principal components. The weighting scheme first estimates the coverage of sequences by fast approximate sequence alignment [185], and the inverse of the coverage assigned to each fragment as the weighting factor. The final distance graph is partitioned using bisection by normalized cuts. Binning clusters are obtained by performing the bisections iteratively.

LikelyBin [186] estimates the genome signatures in the form of Markov models and incorporates them with the *a priori* probability of each fragment which is proportional to the abundance value of the related organism in the metagenome mix. A Markov chain Monte Carlo setting estimates the corresponding probabilities (i.e. genome signatures and population abundance) simultaneously. Consequently, the *a posteriori* probabilities of a fragment for each model indicates the cluster that a fragment belongs to. AbundanceBin [187] is an expectation-maximization algorithm, which uses the Lander-Waterman model [188]. Oligomer frequency estimates are used for the maximization of the *a posteriori* probability of an oligonucleotide coming from a certain species. Once the algorithm converges, the estimated values are used for sequence binning. Tetra [189] is one of the earliest tools used to group the fragments in a metagenome. It uses relative proportions of tetranucleotides with respect to the database samples in DNA contigs and calculates the correlations of pairs as a measure of similarity. In [190], only some of the oligonucleotides, which are believed to carry the phylogeny information, are used for metagenome binning. An approach filtering oligonucleotides which occur with similar frequency between different DNA fragments, as well as the ones with extremely different occurrence statistics improves the binning results. SCIMM [191] is the unsupervised version of the program Phymm. Interpolated Markov models are trained for metagenome fragments and clustered using an expectation maximization algorithm which maximizes the likelihood functions. SCIMM was compared with LikelyBin and CompostBin implementations, and improvement in clustering results were reported. The performance of unsupervised binning algorithms will be compared in chapter 7.

Chapter 6

RAIphy: Phylogenetic Classification of Metagenomics Samples Using Iterative Refinement of Relative Abundance Index Profiles

We have observed that using probabilistic similarity measures instead of absolute metrics can result in better species specific and pervasive characterizations of genome fragments. The RAI measurement, which incorporates several measurements of oligonucleotide abundance based on different Markov assumptions, can possess a pervasive nature with the defined metrics. Although it includes feature extraction of oligonucleotide abundance vectors, which is not pervasive in Euclidian space, it remains sufficiently specific even for short genome fragments using

RAI scores.

In this section, we incorporate this novel genome signature in a semi-supervised metagenome binning algorithm. A given random genome fragment is given a membership score with respect to a taxon by adding up the index values in the RAI model for the taxon for each observed k-mer in the fragment. The fragment is assigned to the taxon that results in the highest score. An iterative process consisting of classifying the fragments from a mixture using the current RAI models then updating the RAI models based on the resulting clusters is used to improve the classification accuracy. As the initial RAI seeds, RAIphy uses models estimated from genomes currently available in the RefSeq database, and thus RAIphy can be categorized as a semi-supervised method. RAIphy has been implemented as a simple, compact standalone desktop application, which is fast compared to similarity-search-based applications. While achieving competitive binning accuracies for the DNA sequencing read length range (100-1000 bp), the method also performs accurately for longer environmental contigs.

6.1 Classification Approach

6.1.1 Classification Metric

To assign a genomic fragment, F, from an unknown source to a taxonomic unit, we first compute the relative frequencies of occurrence for each k-mer from the fragment. For each candidate taxonomic unit, we then obtain a membership score by computing the weighted sum of the components of the RAI profile of the taxonomic unit where the weighting is the corresponding k-mer frequency of occurrence for the fragment F.

Given an RAI model belonging to the taxon, G_i, and an unknown genome fragment, F, the membership score, $E_F[rai^{G_i}]$, is given as:

$$E_F[rai^{G_j}] = \sum_{\mathbf{x}} f_F(x_1, x_2, \ldots, x_k) rai^{G_j}(x_1, x_2, \ldots, x_k), \tag{6.1}$$

where $f_F(x_1, x_2, \ldots, x_k)$ is the frequency of a k-mer in the fragment, F; and rai^{G_j} is calculated using the relative frequency counts of the k-mers observed in the taxon, j. Consider what happens when the statistics of the k-mers of the fragment match the statistics of a taxonomic unit. For a k-mer that occurs often, the frequency of occurrence will be a high and the RAI value of the k-mer for the taxonomic unit will be positive. The more often the k-mer occurs, the larger will be the values of both the RAI and the frequency of occurrence. For k-mers that occur less often than expected, the frequency of occurrence will be low; and the RAI value of the k-mer for the taxonomic unit will be negative. Thus in the sum, the positive RAI values will be weighted by the larger frequencies of occurrence; and the negative values will be weighted with the lower frequencies of occurrence. The opposite will happen when the statistics of the fragments are completely mismatched with the statistics of a taxonomic unit. Therefore, the membership score for the matching taxonomic unit will be higher than the membership score for the mismatched taxonomic unit.

Given the taxa, $J = \{1, 2, \ldots, n\}$, with RAI profiles, $\{rai^{G_1}, rai^{G_2}, \ldots, rai^{G_j}, \ldots, rai^{G_n}\}$, an unknown genome fragment, F, is classified to the taxon, $\hat{j}$, by

$$\hat{j} = \arg\max_j E_F[rai^{G_j}]. \tag{6.2}$$

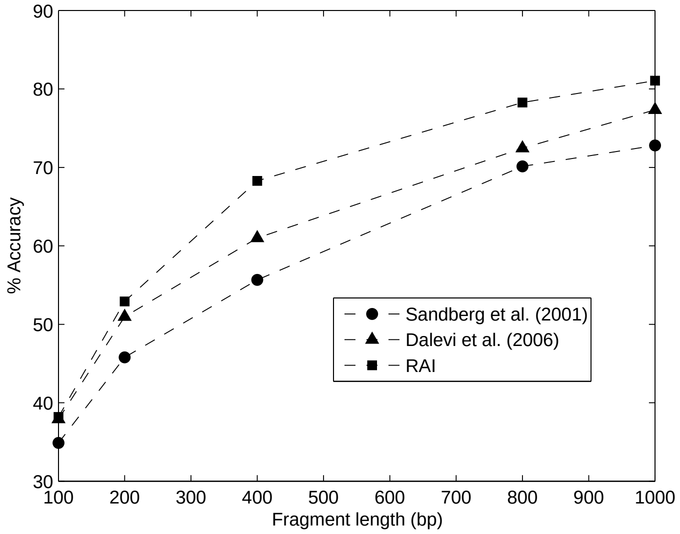

Figure 6.1: The comparison of Relative Abundance Index measure with likelihood measures of oligonucleotide frequencies and Markov models for 100 bp-1000 bp fragment length. Oligomer length of 7 is used.

We compared RAI classification with the detection schemes defined by Sandberg et al. [169] and Dalevi et al. [170] with the same experimental setup used in those studies (Figure 6.1). According to that, random fragments from 28 taxa are classified and the average true positive rations are calculated. RAI was observed to be the best performing method for all fragment lengths in these experiments. Therefore, we have adopted RAI as the compositional detection approach to be used in our metagenomic phylogeny classification.

6.1.2 Iterative Refinement of Genome Models

Metagenomics binning programs are designed for classifying genome fragments of previously unknown species using phylogenetically close genomes. Since the conserved compositional features, or genome signatures, of the unknown species in the mixture are not available, the presumption is that the classification algorithm will assign the fragment to the same or a close clade level for which a model (in this case an RAI profile) is available. While this can be done with some success, there remains significant room for improving the classification accuracy by adaptively updating the models used for detection. The heuristics presented here rely on the fact that we actually possess genomic fragments from the unknown genome in the mixture. Therefore, we use a multistep process in which the first step uses classification, as described above, using the RAI profiles of known species. Once this first classification has been performed, the resulting clusters of fragments can be used to obtain the RAI profiles of the unknown species. Obtaining the genome signatures of these clustered fragments (and subsequently training models over them) results in models that better describe the composition of the unknown genome leading to more accurate classification. Experiments supporting these claims are presented in the Results section.

The refinement procedure consisted of the repetition of two phases. In the first phase, RAI profiles were estimated from genomes of known organisms. Each metagenome fragment was classified by assigning it to the genomes returning the maximum RAI score. In the second phase, the oligonucleotide frequencies and, subsequently, the RAI profiles for each class were recalculated using the collection of fragments assigned to the corresponding class. These two phases were iteratively repeated until a stopping criterion was met. With each refinement, the metagenome fragments were represented with improved RAI profiles. Thus, the average membership scores were expected to increase. When the change in the increase of average membership scores with a refinement became small, we stopped the refinement procedure. Here, the stopping criterion was met if the improvement in the score was less than 1% of the membership score achieved in the previous iteration. The algorithm is quite robust to the stopping threshold; reducing the threshold by several orders of magnitude has no effect on the binning performance. This procedure can be thought of as an expectation maximization algorithm with hard decision of classes [199]. From this point of view, it is similar to a seeded K-means clustering algorithm, with training initial conditions using previously known data [200]. Instead of minimizing the mean Euclidian distance, our objective was to maximize the mean average membership score. The algorithm can be summarized as follows:

```
Classification with iterative refinement:
$N$ Metagenome fragments: $F_j \; j \in \{1, 2, .., N\}$
$M$ RAI profiles: $rai^{G_i} \; i \in \{1, 2, .., M\}$
$M$ taxonomic classes: $G_i \; i \in \{1, 2, .., M\}$


  1. CLASSIFY all $F_j$ using all $rai^{G_i}$
  2. UPDATE all $rai^{G_i}$ using $F_j \in G_i$
  3. BREAK IF $|\frac{AVERAGE\_Membership\_SCORE\_CURRENT}{AVERAGE\_Membership\_SCORE\_PREVIOUS} - 1| < 0.01$
  4. GOTO 1
```

We tested the performance of this algorithm using the same data and experimental design as in [169] (i.e., the same genomes were used for training RAIs, and the same fragments were used for testing). The test fragments in this dataset were short fragments in the range of 100-1000bp. Observing the performance of iterative refinement on short fragments was important because the ratio of false positives is

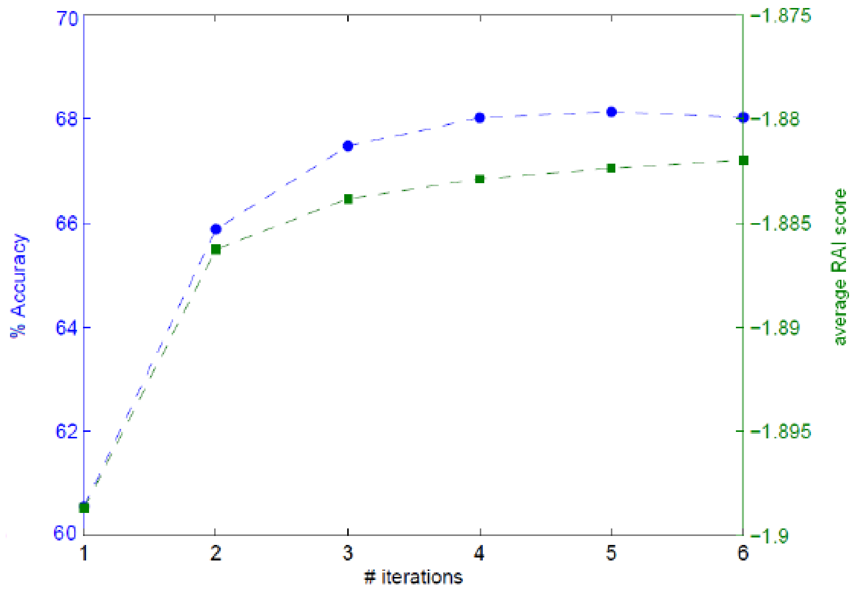

Figure 6.2: The performance increase with iterative refinement is illustrated using the same dataset and experiment setup with [169] for the fragments of length 400 bp. Left y-axis and blue curve: The increase in the percent of correct assignments with iterative refinement. Right y-axis and green curve: The increase and saturation in the average relative abundance index scores.

greater for short fragment lengths, as is the noise introduced by them. Therefore, the task of improving the models in this band was harder. We observed improvement in classification accuracy for all fragment lengths we tested in a small number of iterations (3-6). The increase in accuracy for the fragment length of 400bp is shown in Figure 6.2.

6.1.3 Program Parameters

Since RAIphy was designed as an iterative algorithm, which retrains its models depending on the change in the average membership score, the parameters were kept constant for the whole spectrum of fragment lengths. The oligonucleotide length was fixed at seven. Although it has been shown that longer correlations exist in DNA and that it is possible to exploit longer oligonucleotides for sufficient sequence lengths [202], we observed that the classification accuracy saturates after an oligomer length of seven (Figure 6.3). The binning accuracy increases significantly with the increase in k-mer size to a size of seven. However, increasing the size of the k-mers beyond seven results in negligible accuracy improvement while significantly increasing the computational burden. An RAI profile was updated only if the total length of the fragments assigned to the corresponding class exceeded 25 kbp.

6.2 Results and Discussion

6.2.1 Test Data

In order to be able to conduct controlled experiments, we created synthetic metagenome data using the available genomes in the US National Center for Biotechnology Information (NCBI) RefSeq database [201] as of March 2010. We built our database storing RAI profiles for all 1,146 available genomes. Different chromosomes and plasmids belonging to the same organism were concatenated and treated as a single sequence. These served as the initial seeds in a run of RAIphy. For phylogenetic binning and labeling, we collected the taxonomic information from the NCBI tax-

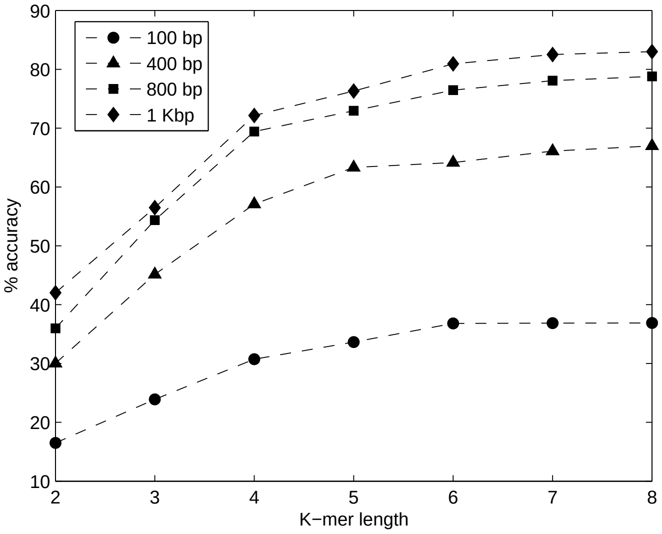

Figure 6.3: The detection accuracy for varying oligomer length using RAI measure in the range of 100 bp-1000 bp fragment length.

onomy database. The data collected was comprised of 609 species, 318 genera, 158 families, 88 orders, 41 classes, and 26 phyla. To test the performance of our program, leave-one-out, cross-validation tests were performed as follows: for every taxonomic unit comprised of at least two subtaxa (e.g., a genus having more than one different species), a test genome was selected; and 3000 test fragments were drawn randomly from each one of those genomes. The RAI profiles were trained over the remaining taxa. The test genome was not used for obtaining the RAI profile. This was done for every genome that was not a single representative of a clade. We repeated each experiment 100 times to assess the first and second order accuracy statistics.

6.2.2 Experiments in Support of the Refinement Process

There are two observations that support the thesis that a refinement process will improve the overall detection performance. First, the genome signatures estimated using the detected portion of a genome should be a good approximation of the signature of the unknown genome. That is to say, we should be able to perform sufficient classification with the models trained from incomplete genomes and even with a collection covering a small percentage of the genome. Although the genome signatures are known to be pervasive, we investigated whether the pervasiveness was sufficient to allow a reasonable estimate of the signature to be extracted from a small fraction of the genome. We repeated the fragment classification experiments in [169] using models trained over various coverage percentages of genomes starting from the entire genome down to only 10% of the genome. Employing the RAI in the manner described above, as shown in Figure 6.4, we observed that there is only a decrease in accuracy of 2-4% in the worst case. This result supports the premise that even with a small collection of fragments in a taxonomic bin after the classification we could train a practically useful model for the unknown organism.

The first experiment demonstrated that it was possible to train a model with a small fraction of the genome that could be obtained through classification of the samples of the microbiome. However, these results assume that the genomic fragments available truly belong to the organism being detected. Taxonomic classification algorithms return significant amounts of false positives. These false positives could conceivably make the algorithm diverge and actually reduce classification accuracy. We conducted a number of experiments to make sure that this would not happen with RAIphy for the metagenomic classification experiments. An example

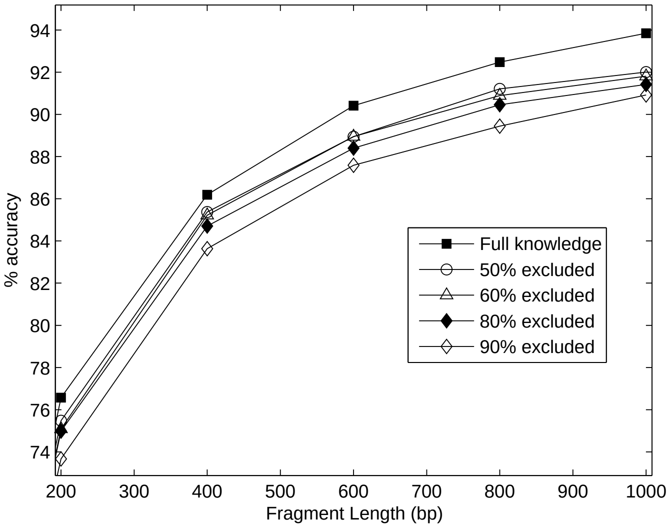

Figure 6.4: Classification accuracy performance with varying available coverage of training genomes. RAI profiles are built using the entire genome and fragments of genomes covering 50%, 40%, 20%, and 10% of the genome. The decrease in the classification performance due to incomplete training data coverage was not significant, and classification capability was conserved.

of the results of such an experiment is shown in Figure 6.2. We had no experiments in which the algorithm diverged.

6.2.3 Classification Performance for Short Fragments

The first set of experiments included testing the accuracy of RAIphy for short fragments in the range of 100-1,000 bp. The experiments were divided into ranges or bands of fragment length, because existing programs operating in different bands have different accuracy scores and properties. For example, TACOA and PhyloPhythia perform poorly for short fragments as mentioned above. On the other hand, similarity-based programs, such as Carma, also perform poorly when the genome of origin is not available. Currently, the only composition-based method that can accurately classify previously unobserved metagenome samples in this range is Phymm. In Figure 6.5, the accuracy (i.e., the percent true positive rate) performance with changing fragment lengths is illustrated. It can be seen that the RAIphy classification performance compares favorably to Phymm for all fragment lengths. In Figure 6.6, RAIphy is compared with PhymmBL, which combines Phymm and BLAST. PhymmBL outperforms RAIphy for shorter fragment lengths at a cost of significantly increased computation time.

6.2.4 Binning Fragments in the Absence of Close Relatives

Even with our contemporary knowledge of microbiology, a great majority of the tree of life is unknown. Therefore, it would not be unexpected to have genome fragments of an unknown clade in a metagenome sample. In this case, a metagenome binning method is desired to assign the fragments of undiscovered genomes to sister

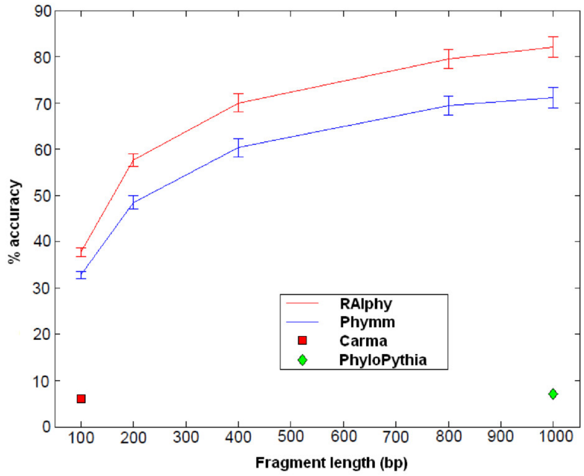

Figure 6.5: Accuracy of RAIphy with short fragment lengths and genus-level prediction, compared with Phymm in the same spectrum. PhyloPythia operates accurately for >1000 bp fragments. Here, its poor performance for short-read range can be observed for 1 Kbp accuracy. Also, Carma searching Pfam domains and protein families for short reads, such as 100 bp fragments, appeared to be performing poorly in accordance with the results reported in [162].

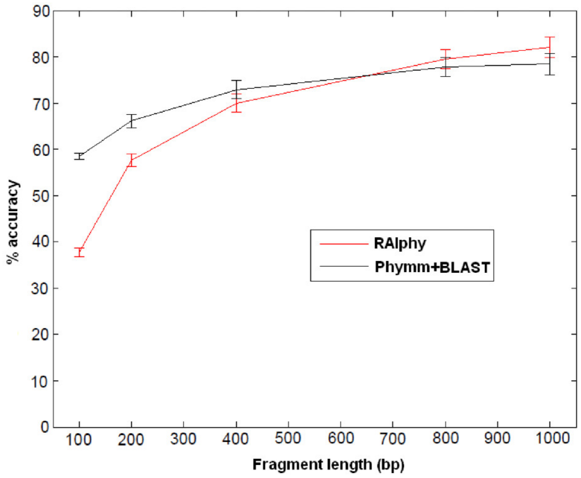

Figure 6.6: Accuracy of RAIphy with short fragment lengths and genus-level prediction, compared with PhymmBL in the same spectrum. For short read length (100 bp-400 bp) fragments, the combination of Phymm and BLAST outperforms RAIphy. However, RAIphy attains higher accuracy for longer fragments.

taxa in the same clade level. To simulate this situation and observe how RAIphy performs in such cases, we tested it with incomplete training data. We repeated the previous experiments with leave-one-out, cross-validation; however, this time, all representatives of the taxonomic group that the test samples belong to were removed from the training data and an assignment to a sister taxon (e.g., a genus from the same family with the unknown genus) was accepted as a correct classification. We performed the tests for the unknown taxa of different clade levels from family to class levels.

The correct classification rate decreased substantially with missing data. RAIphy performed at under 50% accuracy for all clade levels for fragment lengths in the range of 100 bp–1Kbp. In Figure 6.7, the binning performance for RAIphy, Phymm, and BLAST searches is illustrated for a read length of 400 bp and 1 Kbp. While this performance is still superior to other composition-based methods, similarity searches performed using BLAST performed better for short read lengths of 100 bp and 200 bp Figure 6.8, Figure 6.9. For longer fragment length classification, the performance of 800 bp is shown in figure 6.10. The reason why accuracy drops down for the clade level order is the asymmetry in the dataset of currently sequenced microbial sequences.

6.2.5 Classification Performance for Longer Metagenome Fragments

The classification performance for genomic fragments of 800 bp-50 Kbp was also studied. This range is significant because it represents lengths of assembled contigs, while the shorter fragments correspond to single sequencing reads. In taxonomic

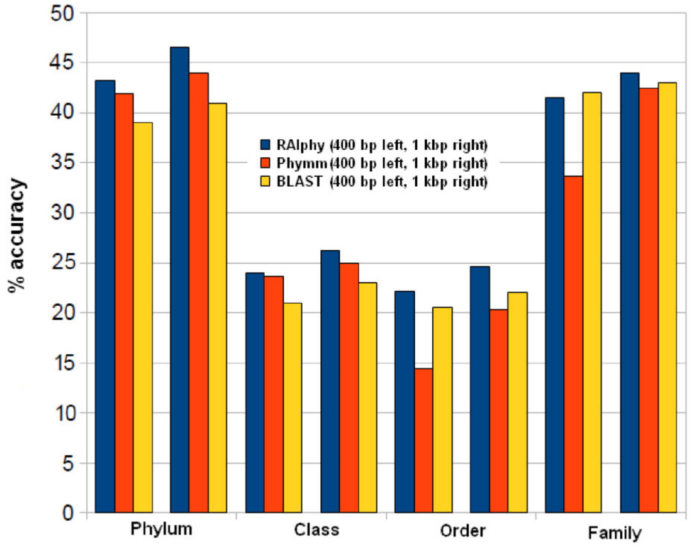

Figure 6.7: Comparison of RAIphy, BLAST and Phymm with incomplete training set for varying clade-levels is shown for 400 bp, 1 Kbp genomic fragments. The accuracy remains under 50% for all methods. RAIphy performs slightly better than Phymm and BLAST for this range.

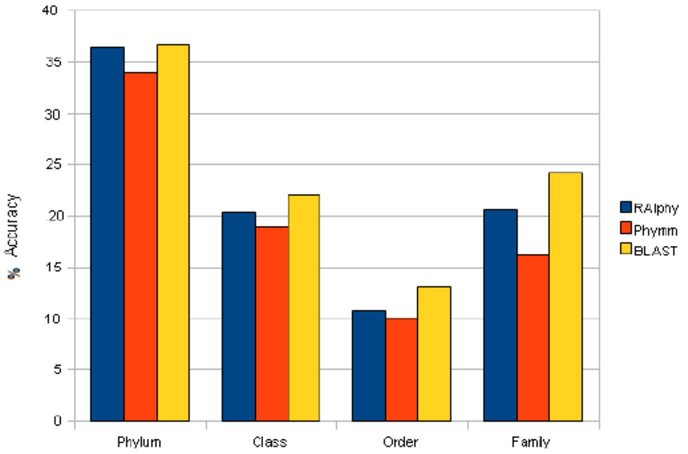

Figure 6.8: Comparison of RAIphy, BLAST and Phymm with incomplete training set for varying clade levels. Fragment length 100 bp.

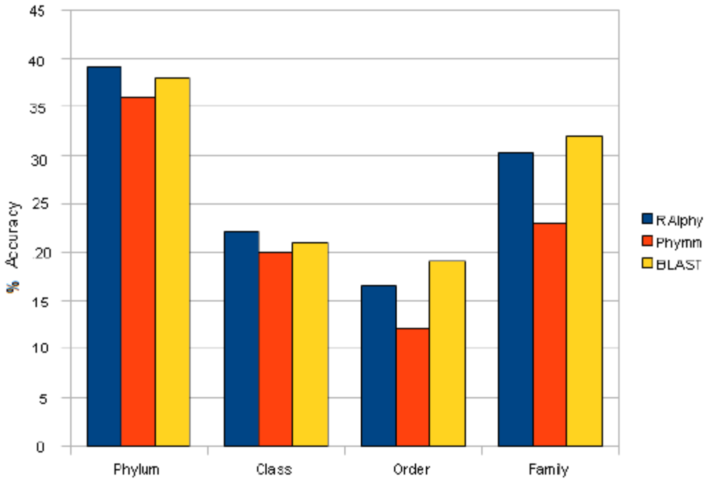

Figure 6.9: Comparison of RAIphy, BLAST and Phymm with incomplete training set for varying clade levels. Fragment length 200 bp.

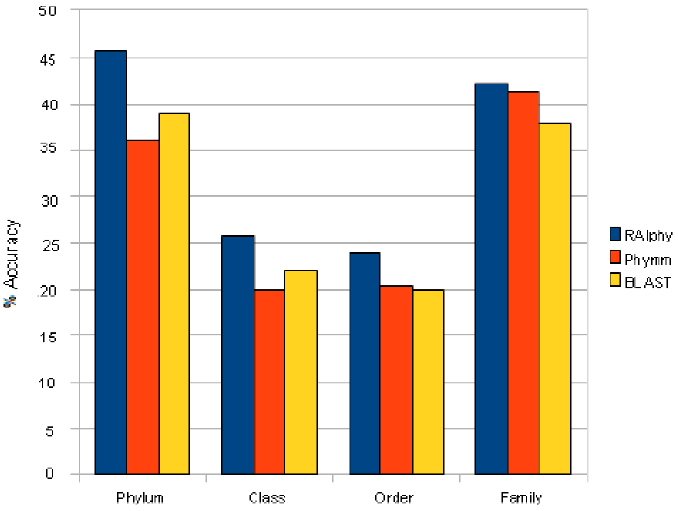

Figure 6.10: Comparison of RAIphy, BLAST and Phymm with incomplete training set for varying clade levels. Fragment length 800 bp.

classification, generation of a smaller number of highly reliable predictions is preferred over predicting the majority of fragments with less reliable labels [172]. When this is the case, genomic fragments with reliable scores can be classified and suspicious fragments left as "unknown." Adopting the accuracy measurement definitions defined by Baldi et al.[203], this kind of regularization yields higher average specificity and lower average sensitivity. The sensitivity for the class i is defined as:

$$Sn_i = \frac{TP_i}{TP_i + FN_i + U_i}, \tag{6.3}$$

where TP_i is the number of samples correctly classified to the class i (*true positives*), FN_i is the number of samples assigned to another class even though they belong to class i (*false negatives*), and U_i is the unclassified number of samples belonging to class i. The specificity for the class i is defined as:

$$Sp_i = \frac{TP_i}{TP_i + FP_i} \tag{6.4}$$

where FP_i is the number of samples assigned to the class i while belonging to another class.

Determining an operating point in the sensitivity-specificity trade-off has been achieved by using different approaches for different methods. In TACOA, the kernel parameters governed the thresholds for classifying samples. In Diaz et al. [172], grid searches were employed to decide the optimal accuracy values and for setting the parameters. PhyloPhytia uses a post-processing one-versus-all SVM classifier to detect the reliable samples and leave the rest "unknown." RAIphy classifies all metagenomic fragments to a taxonomic bin by default. However, RAIphy also allows setting thresholds and operating at different points of the sensitivity-specificity curves. We assigned detection-quality scores to fragments to measure the likelihood of fitting. The quality scores were calculated as the difference between the best average RAI score and the next best score:

$$q(F) = E_F[rai^{G_i}] - E_F[rai^{G_k}] \tag{6.5}$$

where i is the class returning the best RAI score (E_F), and k is the class returning the second highest RAI score. If a fragment fits equally well to more than one model, the quality score turns out to be 0; and if a fragment reflects the characteristics of one class much better than any other class, it receives a high quality score.

Setting percentage thresholds (p) to assign the top p% scored fragments of each class and dropping the labels of the remaining $(100 - p)$ % to “unknown” increased the specificity while reducing the sensitivity. Geometrically speaking, the fragments remaining in an iso-quality hyperboloid were assigned; and the others outside the hyperboloid were determined to be unclassified. Therefore, this thresholding is a tightening of the decision boundary from a hyperplane to a hyperboloid in the feature space.

Figure 6.11, Figure 6.12, Figure 6.13 show the specificity-sensitivity performance obtained from a cross-validation test on the dataset for 800 bp, 1 kbp, and 10 Kbp fragments. Four thousand random fragments were sampled from each test species. The optimized sensitivity and specificity values for TACOA and PhyloPythia were also shown for the same datasets. RAIphy significantly outperformed both algorithms for the given range of fragments and clade levels. An advantage of

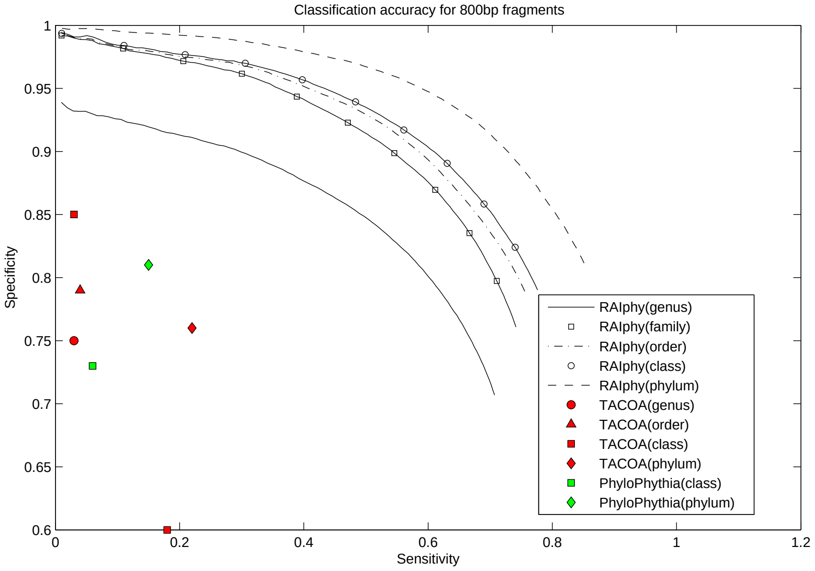

Figure 6.11: Sentitivity-specificity operating characteristics curves for RAIphy determined with 800 bp fragments using the dataset obtained from the RefSeq database. The accuracy values for TACOA and PhyloPythia are also illustrated for the same test data

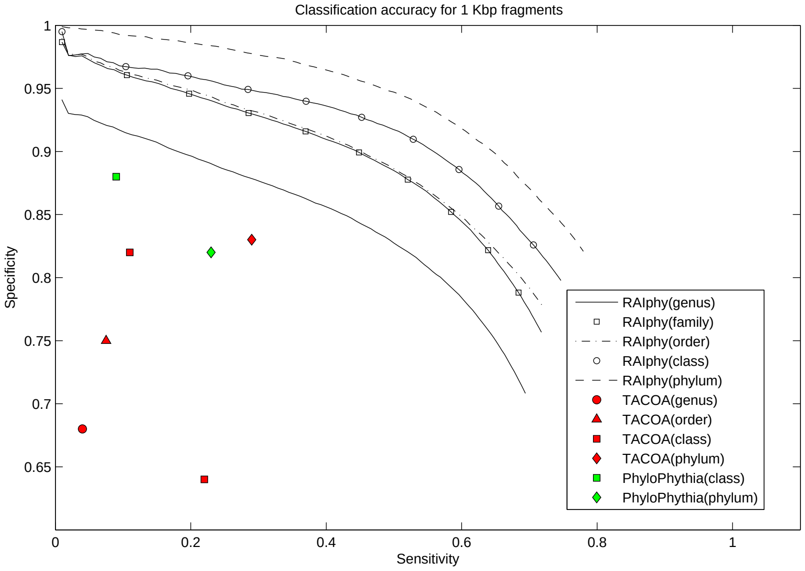

Figure 6.12: Sentitivity-specificity operating characteristics curves for RAIphy determined with 1 Kbp fragments using the dataset obtained from the RefSeq database. The accuracy values for TACOA and PhyloPythia are also illustrated for the same test data.

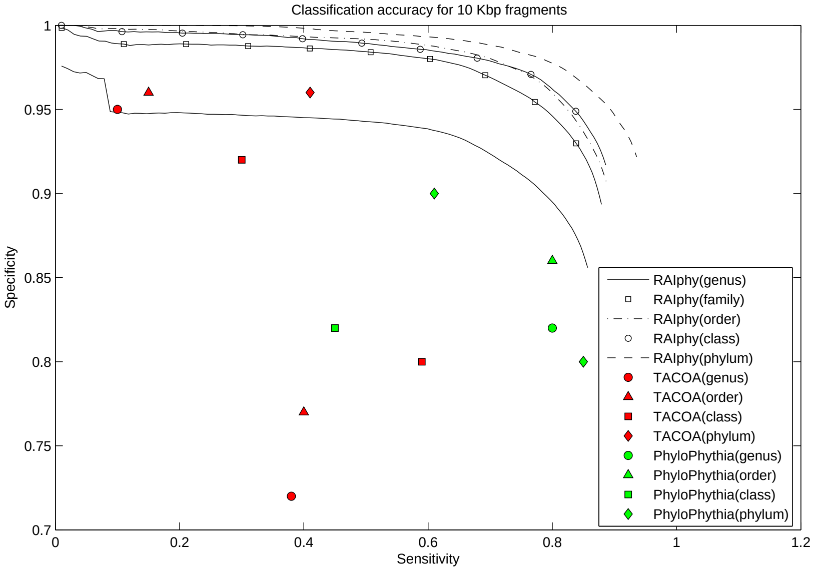

Figure 6.13: Sentitivity-specificity operating characteristics curves for RAIphy determined with 10 Kbp fragments using the dataset obtained from RefSeq database. The accuracy values for TACOA and PhyloPythia are also illustrated for the same test data.

RAIphy, as demonstrated by the sensitivity-specificity performance curves, is that even when samples with low confidence scores are included in the classification, we retain high specificity; and the number of unknown samples decreases and sensitivity values increase, whereas the specificity drop is only around 10-25% for 800 bp and 1 Kbp fragments and around 10-15% for 10 Kbp fragments.

The Specificity performance of RAIphy with the fragment range 800bp-50kbp is provided in Figure 6.14, Figure 6.15, Figure 6.16, Figure 6.17, Figure 6.18 for each taxon at every clade level according to NCBI taxonomy of sequenced genomes.

6.2.6 Performance on Real-Life Metagenomic Data

The RAIphy system was also tested using a real-life dataset. Recognizing the control on real metagenome data is very limited and that true labels of assembled contigs and reads are not entirely known or the labeling is low quality, the experiment was performed on a subset of an Acid Mine Drainage (AMD) metagenome [120]. The AMD sample consisted of a low-diversity community that was dominated by three microbic populations: *Ferroplasma acidarmanus* and *Leptospirillum sp.* groups II and III. Since these organisms exist abundantly in the community, it has been possible to assemble draft genomes for these organisms. Therefore, we can accurately determine which fragment reads belong to these organisms with sequence alignments since fragments originating from the draft genomes align with few mismatches. This allowed us to observe the classification accuracy of our method for a subset of real metagenome data that could be accurately labeled. The phylum-level taxonomy assignments for each of the three genomes are shown in Figure 6.19. *Ferroplasma acidarmanus* belongs to Euryarchaeota phylum of Archaea; 49.6% of

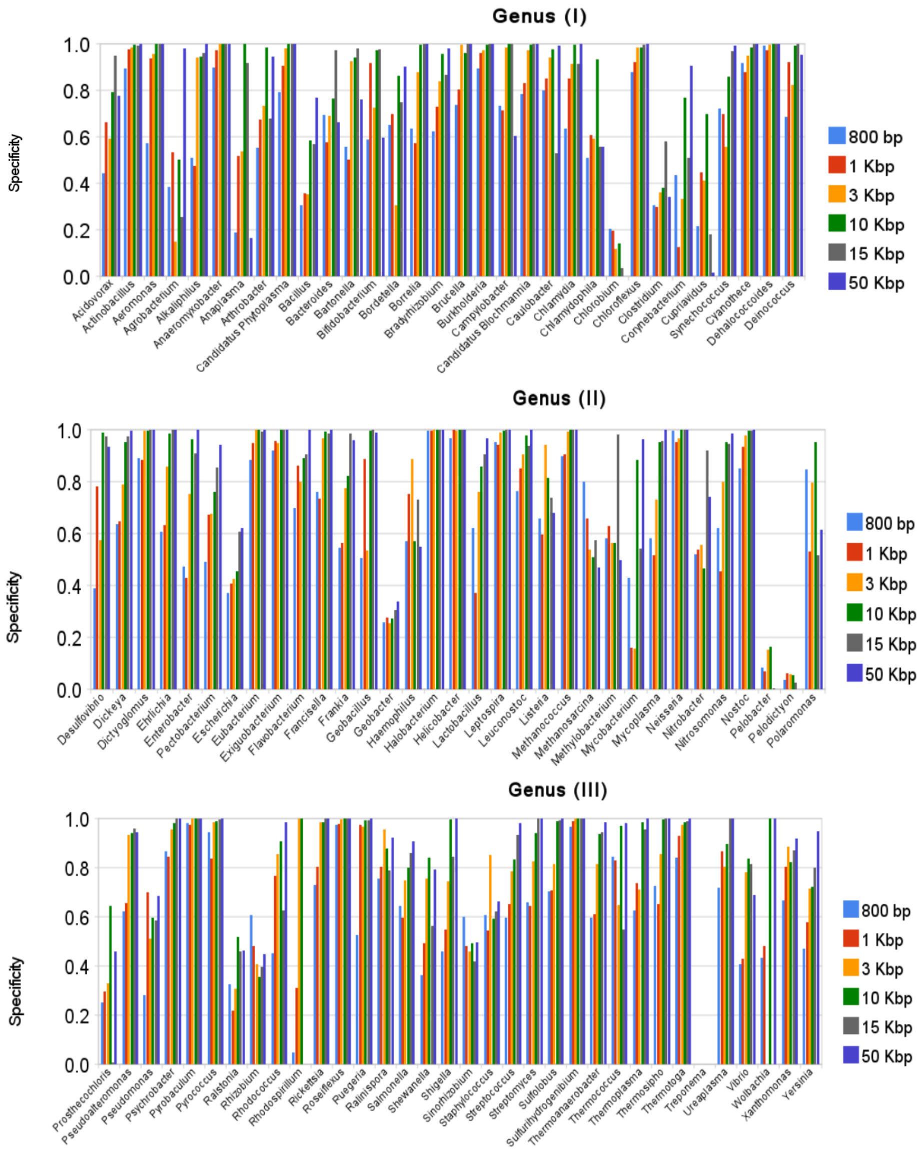

Figure 6.14: Specificity performance of RAIphy in genus level prediction for 99 genera obtained from RefSeq database. Fragment lengts of 800bp, 1Kbp, 3Kbp, 10Kbp, 15Kbp and 50Kbp are illustrated.

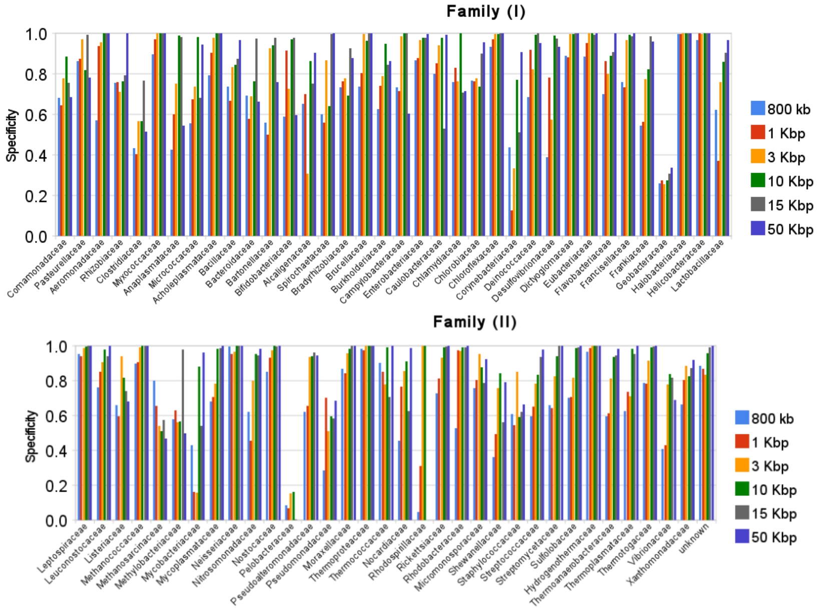

Figure 6.15: Specificity performance of RAIphy in family level prediction for 70 families obtained from RefSeq database. Fragment lengts of 800bp, 1Kbp, 3Kbp, 10Kbp, 15Kbp and 50Kbp are illustrated.

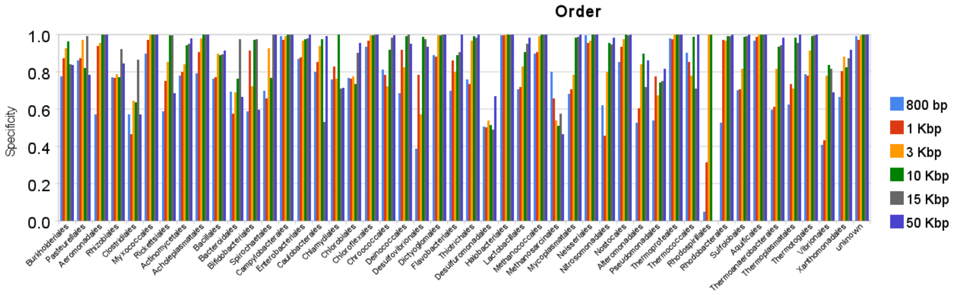

Figure 6.16: Specificity performance of RAIphy in order level prediction for 47 orders obtained from RefSeq database. Fragment lengts of 800bp, 1Kbp, 3Kbp, 10Kbp, 15Kbp and 50Kbp are illustrated.

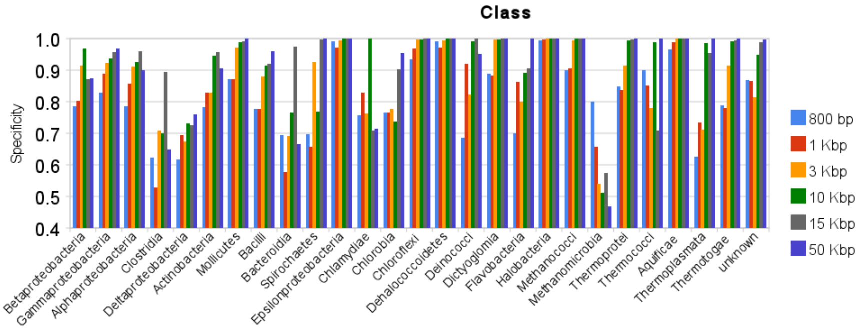

Figure 6.17: Specificity performance of RAIphy in class level prediction for 26 classes obtained from RefSeq database. Fragment lengts of 800bp, 1Kbp, 3Kbp, 10Kbp, 15Kbp and 50Kbp are illustrated.

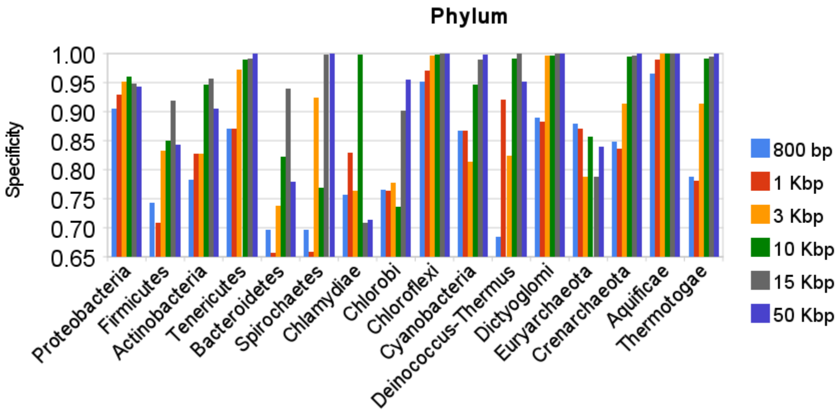

Figure 6.18: Specificity performance of RAIphy in phylum level prediction for 16 phyla obtained from RefSeq database. Fragment lengts of 800bp, 1Kbp, 3Kbp, 10Kbp, 15Kbp and 50Kbp are illustrated.

PHYLUM	Phymm	MEGAN	PhymmBL	RAIphy
Euryarchaeota	41.4%	48.6%	61%	49.6%
Firmicutes	41.9%	18.9%	28.8%	37%
Proteobacteria	8.6%	17.1%	4.9%	5.8%
Bacteroidetes	3.7%	2.2%	2.7%	3.6%
Thermotogae	1.8%	1.2%	<1%	2.1%
Other phyla	2.6%	12%	2.6%	1.9%

Table 6.1: Phylum-level classification of the genome fragments belonging to *Ferroplasma acidarmanus* according to the sequence alignments with the reads and draft of the genome for the taxonomic classification programs Phymm, MEGAN, PhymmBL, and RAIphy. Correctly classified phylum is Euryarchaeota.

PHYLUM	Phymm	MEGAN	PhymmBL	RAIphy
Proteobacteria	80.2%	60.4%	79.6%	87.6%
Chlorobi	6%	2.5%	5.7%	4.9%
Firmicutes	2.3%	10.2%	2.7%	2.1%
Actinobacteria	< 1%	1%	2%	1.3%
Other phyla	2.6%	12%	10%	2.2%

Table 6.2: Phylum-level classification of the genome fragments belonging to *Leptospirillum sp.*group II according to the sequence alignments with the reads and draft of the genome for the taxonomic classification programs Phymm, MEGAN, PhymmBL, and RAIphy.

the fragments were correctly classified, as shown in Figure 6.19-a. This compares with 41.4% for Phymm , 48.6% for MEGAN, and 61% for PhymmBL, as shown in Table 6.1. The similarity scores used as MEGAN input were obtained from nucleotide BLAST with the RefSeq database used as the similarity search set.

Leptospirillum sp. groups II and III are bacteria belonging to the Nitrospirae phylum, which does not exist in the NCBI RefSeq database and, consequently, in our database. The genus *Leptospirillum* was assigned as Deltaprotobacteria [204], which is a class of Protobacteria. Of the fragments putatively determined to be *Leptospirillum sp.* group II reads, 87.6% were assigned to the Protobacteria phylum. For Phymm the true positive percentage was 80.2%, for MEGAN it was

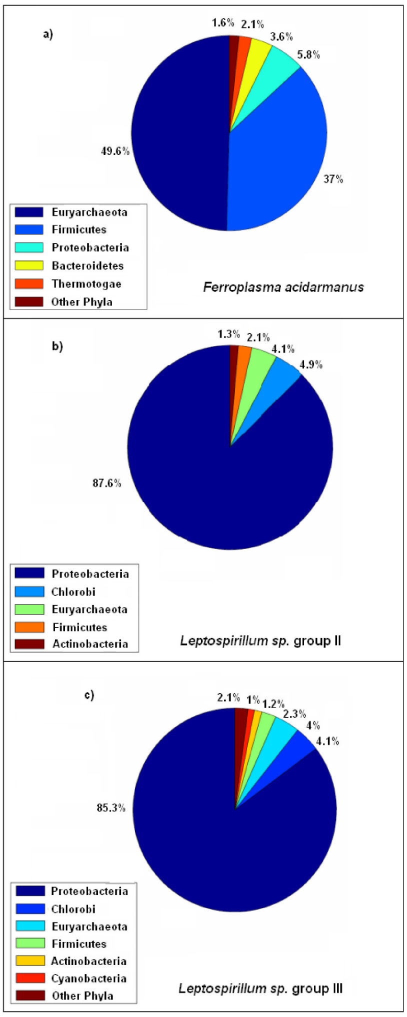

Figure 6.19: Phylum-level classification of the AMD metagenome fragments.

PHYLUM	Phymm	MEGAN	PhymmBL	RAIphy
Proteobacteria	77.3%	62%	76.9%	85.3%
Chlorobi	3.9%	1.7%	3.3%	4.1%
Euryarchaeota	8.4%	4.9%	7.7%	4%
Firmicutes	2.7%	6.8%	2.9%	2.3%
Actinobacteria	2%	12.7%	3.3%	1.2%
Cyanobacteria	1.1%	3.8%	1.3%	1%
Other phyla	4.6%	8.1%	4.6%	2.1%

Table 6.3: Phylum-level classification of the genome fragments belonging to *Leptospirillum sp.*group III according to the sequence alignments with the reads and draft of the genome for the taxonomic classification programs Phymm, MEGAN, PhymmBL, and RAIphy.

60.4%, while for PhymmBL it was 79.6%. Finally for *Leptospirillum sp.* group III fragments, the true positive rate for RAIphy was 85.3%. This compares to 77.3% for Phymm, 62% for MEGAN, and 76.9% for PhymmBL. This is a significant improvement in classification performance.

6.3 Conclusions

A metagenome binning method that exploits inherent features of genomic signatures with a novel measure called RAI and a novel classification metric is proposed. Our simulations used a large genomic fragment length range from 100 bp to 50 Kbp. This range covers the length of average metagenome assembly contigs and the length of sequencing reads with the current sequencing technology. The simulations resulted in classification accuracy ranging between 38-97% at the deepest clade level (genus). Using RAI scores, the optimal performance was obtained using relatively longer oligonucleotides (7-mers) than methods using Euclidian distance and correlation-based scores utilizing shorter k-mer statistics. We attributed a part of the improvement in classification accuracy to being able to use longer oligonucleotide statistics, which include additional information on the DNA k-mer distribution. Moreover, with the availability of RAI profile updates using the predicted DNA sequences, we have defined an iterative classification method that improves the classification accuracy. We believe the improvement is due to the fact that genome signatures are pervasive, and genome models can be approximated without requiring the availability of complete genomes. Therefore, a small set of genome fragments was sufficient to update the initial genome models. In our case, a set of fragments forming 25 Kbp of nonoverlapping genomic sequence was sufficient to increase the classification accuracy in the next iteration.

In addition to the experiments performed on synthetic metagenomics data, we tested RAIphy with well-studied, real-life metagenome AMD sample reads. RAIphy outperformed the composition-based Phymm and nucleotide BLAST search-based MEGAN on the binning task. PhymmBL, which uses a composite method consisting of Phymm and BLAST, did better than RAIphy in one of the three tasks and worse in the other two. PhymmBL took substantially longer to complete the tasks than Phymm or RAIphy (around 5 fold longer).

The running time of RAIphy scales linearly with the average fragment length and the number of fragments in the metagenome sample. In our experiments, it took less than 4 hours to bin the AMD metagenome with the most comprehensive search models that contained all 1,146 genomic sequences of the (NCBI) RefSeq database on a standard desktop computer with a 2.19 GHz CPU. Processing of the same dataset with similarity-search-based binning programs, such as CARMA and MEGAN (run with blastn), and even with phylotyping pipelines AMPHORA and MLTreeMap, requires $>$ 24 hours. PhymmBL took around 464 hours to process the dataset. Using genus level RAI profiles, the current version of RAIphy can bin 1.5 Gbp of genomic sequences with 400 bp average read length in 24 hours. This amount of data is achievable with next generation, high-throughput sequencing; and RAIphy appears to satisfy a computational need for fast and accurate metagenome binning. RAIphy uses a moderate amount of memory (304 MB with species-level training loaded and 47 MB with genus-level training loaded) in its runtime.

We have implemented RAIphy as an open-source desktop application supported with a simple graphical user interface. While the default is for all the RAI profiles of the RefSeq database in the species and genus level to be used as database files, there is also an option to create custom databases if a set of training sequences are provided. Since the program performs with a satisfactory accuracy both for read-length and assembly-length DNA fragments, it can be utilized either as a preprocessing stage in a metagenomics pipeline to improve the assembly procedure or as the binning procedure for the assembled contigs.

We have observed that the accuracy falls to below 50% when sister taxa of the unknown fragments are not close relatives. This appears to be a universal problem that is also observed with other binning methods. For the metagenome samples of undiscovered microbes, it might be a safe strategy to sacrifice prediction resolution and bin the sequences to higher taxonomic units, such as phylum or class, or sacrifice specificity by selecting best hits and leaving suspicious assignments "unknown." RAIphy outputs assignments at all taxonomic levels as well as providing a thresholding option to select the best hits. Another universal problem, which RAIphy also suffers from, is the classification of horizontally transferred regions in procaryotes. Since recently transferred regions differ in composition, predictions of those regions result in false binning.

Chapter 7

Unsupervised Binning of Metagenome Samples

7.1 Metagenome Assembly Problem

Metagenomics, as a newborn science, has provided valuable achievements for areas such as clinical microbiology, virology, evolutionary biology as well as medicine and industry. The promises of this emerging field might be broader than what has been achieved in its first decade. However, to explore the further opportunities, many open research problems have to be addressed. Perhaps one of the most important issues to be considered is the problem of taxonomic assignment of environmental samples. In this chapter, we introduce a novel paradigm addressing the taxonomic assignment problem, using the concept of genome signatures. A metagenome binning application developed in this direction provides significant improvement over conventional approaches.

The current paradigm of computational metagenome analysis breaks down the taxonomic annotation process into two subsequent major phases: metagenome assembly and binning. A general analysis strategy in a metagenome project is the assembly of obtain longer contigs prior to binning fragments in OTU's. Direct taxonomic assignment of sequence reads are generally avoided because both for similarity search and composition based binning methods, short fragment lengths result in poor binning [205]. Considering that the second generation sequencing technology outputs sequence reads around the range of a few hundred base pairs, an assembly phase before taxonomic assignment appears to be an appropriate strategy.

The unavailability of metagenome binning prior to the fragment assembly process leads to an unusual assembly problem that includes the shotgun multiple assembly of various genomes. Conventional genome assemblers are designed for single genomes; however, existence of multiple organisms in a metagenome introduces a number of problems that reduce the quality of contig assemblies. These problems can be reviewed as follows.

7.1.1 Interspecies Chimeras

Since genome assembly programs are designed to yield single genomes, aggressive attempts to cross-assemble different genomes can result in the creation of chimeric sequences [206] due to the existence of homologous regions in different genomes. As this problem does not exist for assembling single genomes, there is no mechanism to avoid this in single-species assemblers.

7.1.2 Non-homogeneous Coverage Distribution

Because different organisms vary in abundance based on the population dynamics of the ecosystem, single nucleotide coverage appears to be different for different organisms. Although this could be a discriminative feature for distinguishing different organisms in an environmental sample, it can become a disadvantage with conventional genome assemblers. For example, the Celera Assembler treats regions with atypical coverage as repeat regions and avoids assembling them [207] as homogenous coverage is expected in single genome data and atypical coverage is assumed to be because of overlapping repeat regions.

7.1.3 Large Amount of Sequence Data

Most genome assemblers are designed for Sanger sequencing data with longer fragments and less coverage depth. One class of assemblers attempts to solve a Hamiltonian path problem where each read is a vertex of a graph and edges are the overlaps between those reads. With shorter fragments and greater coverage obtained by new generation sequencing, the number of vertices increase significantly especially with Metagenome data. This increases the computational complexity to a point where the approach is no longer feasible [208,209].

7.1.4 Existence of Different Strains of a Species with a Number of Polymorphisms

Single genome assemblers which use Eulerian graph solutions generate De Brujin graphs to solve the assembly problem. Errors in reads expand the graphs and increase the computational complexity so error correction mechanisms are used as preprocessing steps to reduce the size of the De Brujin graphs. However, in addition to sequencing errors which are typically no more than 3% in a fragment, polymorphisms due to coexistence of different strains and individuals are introduced in environmental samples which increases the edit distances of reads from the consensus sequence and expands the graphs. This can affect assembly performance of Eulerian-path methods. In fact, the error-correction mechanism of Eulerian-path methods cannot perform efficiently for high variation data [210] which makes the assembly of metagenomes nearly impossible in that framework. Moreover prophages inserted in different locations within the same species, or genomic rearrangements make a single consensus sequence (which is the goal of a single-genome assembler) unachievable [211]. This appears to be another reason for the highly fragmented metagenome assemblies where the fragments are mostly the consistent portions of different strains shared in all genomes of a species [212].

Genome assembly methods can be divided in two phases: detecting overlaps and joining the reads. Pairwise similarity searches of reads using common k-mers or sequence alignments are mainly used for detecting the overlaps. The main focus in these algorithms is on connecting the overlapping reads while avoiding repeats. The algorithms mainly differ in how they solve the path-finding problems. In analyzing metagenomic data, the attention should be on avoiding interspecies coassembly instead of avoiding intraspecies coassembly of repeats. It should be considered that the performances of the assembly and binning phases (and, consequently, the overall performance of the whole analysis process) not only depend on each other but have many elements in common. We introduce a taxonomy assignment approach to the metagenome analysis problem that would jointly perform taxonomical binning and emulate contig assembly in a more careful and specified fashion.

7.2 Unsupervised RAIphy

Current approaches in the metagenome analysis of microbiomes consist of independent processes of fragment assembly and binning. However, communication between these two processes can contribute significantly to the accuracy of both processes since they each produce important information for the other. Unsupervised RAIphy is a method which intimately combines concepts from binning and assembly to produce a final taxonomic classification. It basically employs the overlap process of fragment assembly to grow longer pseudocontigs and an unsupervised binning procedure is performed over the pseudocontigs. Both processes use the concept of genome signatures with relative distance measurements, in order to detect sequencing read overlaps and contig clustering, respectively. Moreover, intermediate procedures to mitigate the problems associated with metagenome asembly are developed. Particularly, avoidance of chimeric sequences is addressed. Two intermediate procedures between pseudoassembly and pseudocontig clustering attempt to segment contigs with interspecies chimeras. First, we introduce the consecutive steps of the algorithm.

7.2.1 Fragment Walk

The first step of unsupervised RAIphy attempts to grow pseudocontigs in parallel by joining together reads which have the same compositional structure and display some overlap in a greedy fashion. The aim in this procedure is to group fragments coming from the same region of a genome. This group of reads will cover a longer part of the genome, resulting in a better estimation of the genome signature to be used in binning.

The corresponding procedure of joint pseudoassembly/binning method is called a fragment walk. Each new read is either used as the seed for a new fragment or it extends an existing fragment in a depending on whether there is a fragment to which the read is close. A fragment walk might result in chimeric sequences as the reads are from multiple organisms some of which may be evolutionarily related and, therefore, have similar compositional structure.

In Figure 7.1, the layout of simulated sequencing reads from a small genome region with the FLX 454 sequencing technology with 10X coverage is shown. The data are generated on *E. coli* genome using MetaSim simulator with 400 bp average read length. A similar layout using random fragmentation with uniform distribution of cut points is shown in Figure 7.2.

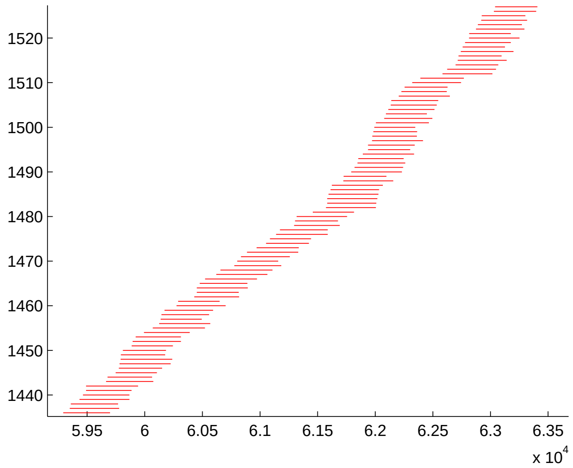

Figure 7.1: FLX 454 400 bp fragment reads layout.

It can be seen that FLX 454 reads exhibit similar character with uniform distri-

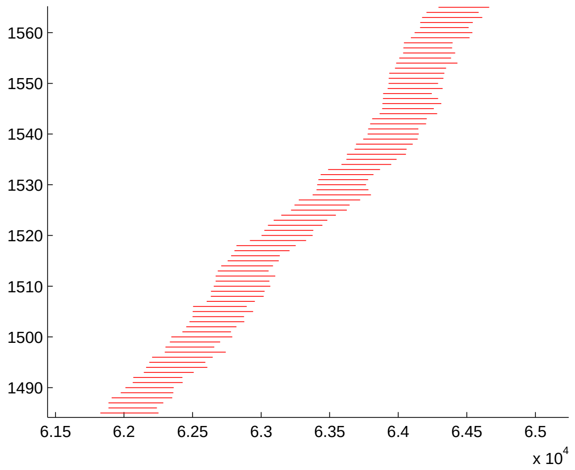

Figure 7.2: Uniform 400 bp fragment reads layout.

bution of sequencing reads. Therefore, a fair overlap distribution depending on the coverage of the corresponding genome is expected. The overlaps of adjacent reads imply similar sequence characterizations since they share the sequence of the overlapping region. Moreover, similar signature distance values are expected within a genome since the read distribution is uniform-like. Use of genome signatures for overlap detection also has the advantage that picking reads from other genomes is unlikely due to the species specific character of genome signatures.

Given a genome signature characterization $G(f_i)$ where f_i is the i^{th} sequence read in a set of N reads ($i \in \{1, 2, \ldots, N\}$), the closest overlapping read of read i is determined to be

$$\overline{j} = arg \min_{j} D(G(f_i), (f_j)) \tag{7.1}$$

where $D(.,.)$ is a distance metric for fragment characterizations. This procedure is iteratively repeated by replacement until a stopping criterion is met or until no fragment reads are left.

In order to provide evidence for the validity of this approach, we generated a synthetic dataset with simulated reads from six bacterial species. In our experiments, Markov models with likelihood function metric was the best performing genome signature. In Figure 7.3 the fragment walk performance of this signature is compared with hexamer frequency vectors with Euclidean distances.

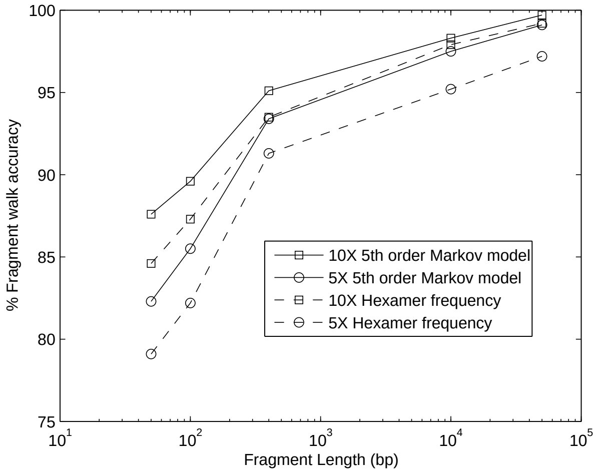

Figure 7.3: The fragment walk performance comparison of 5^{th} order Markov model with likelihood function metric and hexamer frequencies with Euclidean distance.

The simple greedy method described above was used to generate contigs for different read lengths and different coverages. The results are shown in Table 7.1. For the same dataset, using 454 sequencing simulations with 10X average coverage and the stopping heuristics - stop walk when the minimum distance is greater than 5 times the average minimum distance in the previous history - a simple greedy pseudocontig generation approach resulted in approximately 30 Kbp long contigs. Having long contigs of this size is highly advantageous for taxonomical clustering.

	Read Length (bp)				
Coverage	50	100	400	10000	50000
3X	80.5	82.4	89.2	96.2	97.8
5X	82.3	85.5	93.4	97.5	99.1
10X	87.6	89.6	95.1	98.3	99.7

Table 7.1: Fragment walk accuracy for the dataset *Escherichia coli, Pseudomonas putida, Thermofilum pendens, Pyrobaculum aerophilum, Bacillus anthracis* and *Bacillus subtilis*. The relative abundance ratios of the organisms are 1:1:1:1:2:14. Accuracies are calculated for read lengths of 50, 100 (Illumina, Helicos, SOLiD sequencing), 400 (454 sequencing), and 10 Kbp, 50 Kbp (nanopore sequencing) where average coverage values of 3X, 5X and 10X are simulated.

7.2.2 Segmenting Chimeric Sequences

The fragment walk procedure generates groups of reads which are expected to be belonging to the same part of a genome. Since these reads are short DNA fragments, the similarity metric based on Markov model characterization might have a relatively high ratio of failing to detect samples from the same genome. This could be because of the weakness of the signature at that fragment length level as well as the homogenous regions existing in different genomes. Therefore, the problem of interspecies chimeras which is common to metagenome assembly has to be addressed.

Coverage Based Segmentation

In order to deal with the problem of chimeric sequences, in the second step, unsupervised RAIphy attempts to segment the sequence generated by the fragment walk using a segmentation algorithm premised on the non-uniform abundance of different genomes. For the postprocessing of the contigs generated we intend to utilize the statistical differences between different organisms due to their distinct relative abundance values in the community. Abundant species are sampled more frequently from environment, and thus they have greater number of sequencing reads. Due to the Lander-Waterman equation, greater number of reads mean higher sequence coverage and longer overlaps between the adjacent reads. As a result, the distance scores obtained during the fragment walk are varies according to the relative abundance of species in the community. We can see this phenomenon in Figure 7.4 where the first 50 sorted distances averaged over 20 random fragments selected from 454 reads of length 400 bp with 3X, 10X, and 20X coverages. Unsupervised RAIphy stores the best 10 distance scores belonging to a read calculated during the fragment walk phase.

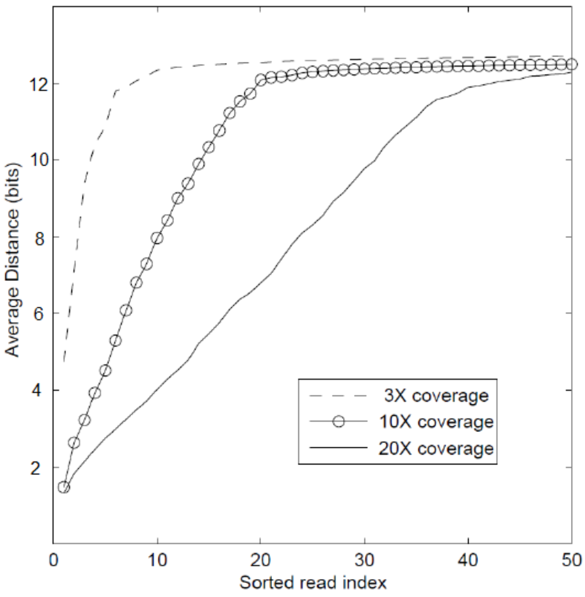

Figure 7.4: Sorted distances (the first 50 components) averaged over 20 random fragments selected from 454 reads of length 400 bp with 3X, 10X, and 20X coverages

Assigning an average-smallest distances score to each read in a pseudocontig,

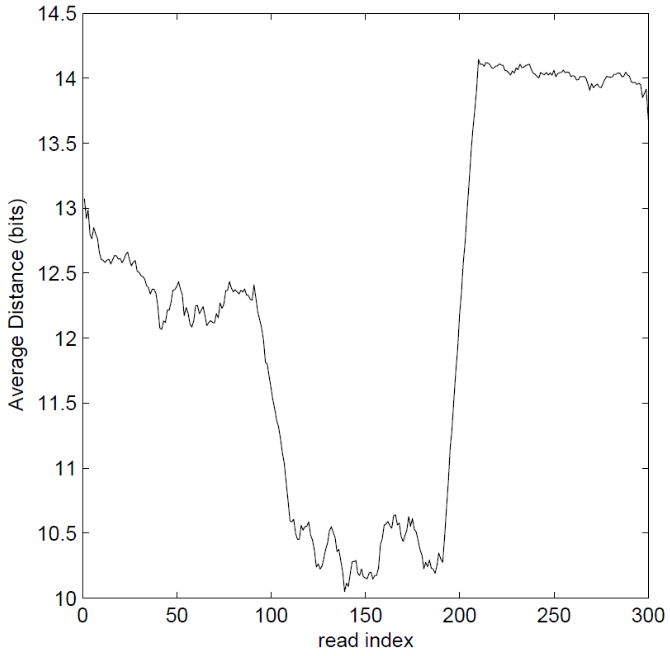

Figure 7.5: scores for an artificial chimeric contig with the reads for the first hundred bases from a genome with 10x coverage, bases 101-200 for a genome with 20X coverage, and bases 201-300 for a genome with 3X coverage.

a numerical signal is generated such as the one plotted in Figure 7.5. The scores derived from an artificial chimeric contig consisting of 3 regions from different genomes with coverage values 3X, 10X and 20X is illustrated. Each read is represented by the average of first 10 smallest distances. It can be seen that high coverage regions have smaller average scores and that the average score increases with decreasing coverage values. Unsupervised RAIphy processes the signal with a simple method: a change in regime determined if 20 upstream values in the signal are significantly different than the 20 downstream values. Empirically, this was decided as the student's t-test values, and when ever a t score $t > 1.5$ is met, the boundary is detected as a segmentation point of chimeras.

Composition Based Segmentation

Interspecies chimeras problem created during the fragment walk procedure can be mitigated when the populations of chimeric organisms are different from each other. However, species with close relative abundance in the community are more difficult to distinguish by this method. As a complementary second phase, unsupervised RAIphy employs a procedure that makes use of genome signatures in an information theoretic framework. An approach such as entropic segmentation is applicable in case the species diversity is not discriminative. We determined Jensen-Shannon divergence [213] as the segmentation method for composition based segmentation of unsupervised RAIphy. According to this scheme a moving boundary bisects a genome fragment f in two parts f_a and f_b. The divergence between these two subfragments is

$$D_{JS}(f_a, f_b) = H(f) - \frac{|f_a|}{|f|}H(f_a) - \frac{|f_b|}{|f|}H(f_b) \tag{7.2}$$

where the entropies are estimated using relative trinucleotide frequencies as distribution estimates. The divergence scores are calculated for a moving boundary along a DNA fragment and the boundary position resulting in the highest divergence score is detected as the segmentation boundary. As an empirical threshold, we decided to segment the sequence if the peak divergence value is greater than 2 times the average divergence along the fragment.

In Figure 7.6,the segmentation algorithm is run on an artificial chimeric contig with constant depth coverage. Unsupervised RAIphy employs JS segmentation iteratively until no segmentation is required according to the segmentation threshold.

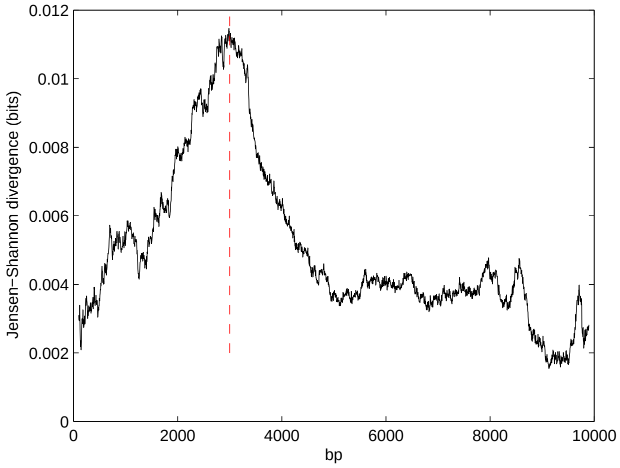

Figure 7.6: Jensen-Shannon divergence values for a moving boundary in a chimeric sequence. The first 3000 bp is assembled from *E. coli* (GC-content: 0.51) and the last 7000 bp is assambled from *Y. Pestis* (GC-content: 0.48). Distributions of 3-mers were used. The maximum JS divergence value correctly locates the boundary of species transition.

Even in the absence of chimeras the fragment walk may result in the joining together of noncontiguous sections of the genome because of the pervasiveness of compositional structure. Furthermore, the fragment walk ignores the problem of repeats which will also result in a misassembly. For these reasons we call the resulting assembly a pseudocontig instead of a contig. However, these issues are not to be addressed by unsupervised RAIphy, because the primary goal is to group the taxonomically related metagenome data together.

7.3 Taxonomic Clustering

Taxonomic clustering of pseudocontigs generated in previous steps is be performed by an unsupervised clustering algorithm. According to this scheme, every contig is represented by the compositional features. Relative abundance indices are trained for each cluster, and RAI metric is employed to classify the pseudocontigs. A simulated annealing technique is used to avoid local minima problem.

7.3.1 Estimating the RAI's of Taxa Existing in the Metagenome

When models for the source genomes are available *a priori*, as the case in supervised binning, the problem is a *detection* problem in which we detect the classes of fragments given the models. When models for the genomes are not available we can treat the problem as an *estimation* problem in which the models describing the given fragments are to be estimated. Since we quantify the goodness of a model with the RAI score needed with respect to the model for a given fragment, the objective function we need to maximize is the total relative abundance index:

$$\widehat{\mathcal{M}} = arg \max_{\mathcal{M}} rai(\mathcal{M}|C, X, n) \tag{7.3}$$

where $\mathcal{M}$ is the set of RAI models, C is the class assignments X is the set of fragments (or pseudocontigs) and n is the number of models. In other words, we search for the model set which would most parsimoniously represent the fragment collection. The solution of this optimization requires partitioning the set of fragment into classes. While this parsing problem is known to be NP-hard [200], efficient approximate solutions exist. We use the Generalized Lloyd Algorithm [198] also known as the Linde-Buzo-Gray (LBG) procedure [214] for this optimization.The classification procedure can be abstracted as follows:

```
PROCEDURE::LBG

  1. initialize models
  2. perform RAI detection on the fragment set
  3. update models training over the concatenation of fragments
     in each class
  4. if the change in the objective is not significant, TERMINATE
  5. GOTO 2
```

The LBG algorithm converges to a local minimum with a solution which is dependent on the initial conditions rather than the global minimum. In our case, we initiate models training over random pseudocontigs from the set generated in previous steps of the algorithm. If two initial fragments picked are from the same taxon, the corresponding models will tend to describe the fragments from the same taxon and multiple genomes will be described by single models. This unbalanced distribution causes the algorithm to settle down to a local minimum. In order to mitigate this problem, we modify the LBG algorithm in two ways:

1. Initializing models by picking dissimilar fragments
2. Simulated annealing based on disturbing “abnormal” models

Initializing Models by Picking Dissimilar Fragments

In order to start with initial models estimated from the fragments of different taxa, we iteratively split the initial models training them over the most dissimilar fragments available. We begin by picking a random fragment and train a model using it. In each iteration a new model is added by training it over the fragment which has the minimum RAI score in the set. We expect that this fragment will belong to an unmodeled genome and the new model will desctribe a new genome in the mixture.

```
PROCEDURE::initialize models

  1. train a model over a random fragment
  2. perform RAI detection on the fragment set
  3. find the fragment with minimum RAI score, train a new model
     using it
  4. if the number of models is reached, TERMINATE
  5. GOTO 2
```

Simulated Annealing Based on Disturbing Abnormal Models

Simulated annealing can be employed as a refinement to LBG by disturbing the final solution and trying to perturb it from a local minimum. In our case, we can disturb bad models, forcing them to evolve to better estimations. That is to say if a single genome is described by multiple models or vice versa the corresponding models should be significantly different from models which uniquely describe single genomes. Moreover, if we can define functions detecting this kind of abnormality, it is possible to define procedures for disturbing final solutions to force the optimization converge to global maximum. Here we define statistics which can be used to differentiate good models from bad models.

Model Quality Based on Population Density

In order find some characteristics of good models, we first start with the hypothesis that a genome fragment is likely to have a large RAI score calculated under the model estimated for that taxon, i.e. given the taxon model, fragments with higher RAI scores are observed with higher probability. Now we define a function which gives a measure of deviation from the taxon model: Assume Y is the pseudocontig that is used as the training sequence for the estimation of the taxon model and x is any random fragment from this taxon. The RAI score distance of these two sequences can be defined as the difference in the amount of information in bits to describe the sequences x and Y:

$$\begin{aligned} D_{RAI}(x, Y) &= |rai(x|Y) - rai(Y)| \qquad (7.4) \\ &= \sum |f(x_n, x_{n-1}, \ldots, x_{n-k}) - f(Y_n, Y_{n-1}, \ldots, Y_{n-k})| rai(Y_n, Y_{n-1}, \ldots, Y_{n-k}) \end{aligned}$$

When the fragment x has the same oligonucleotide frequencies with the training sequence Y, the description divergence is zero and it is some positive value for diverging oligonucleotide distribution. For a good model, the RAI distance will have a probability distribution function which is greater for large values with a large kurtosis. Note that this is also associated with having large total RAI score. However, this might not be true for a weak taxon model. We measure model fitness with a function of RAI distance:

$$\mathfrak{F} = \frac{R_{0.5}}{R_{0.1}} \qquad (7.5)$$

where

$$P(D_{RAI}(x,Y) < R_{0.5}) = \int\limits_{r<R_{0.5}} p(D_{RAI}(x,Y) = r)\, dr = 0.5 \tag{7.6}$$

and

$$P(D_{DL}(x,Y) < R_{0.1}) = \int\limits_{r<R_{0.1}} p(D_{RAI}(x,Y) = r)\, dr = 0.1 \tag{7.7}$$

We compare the radii that the probability of observing a genome fragment with RAI score greater than them are 0.5 and 0.1 respectively. So since for the good models the probability density function reaches its largest values for large RAI scores, it is expected that the CDF reaches 0.1 in a small radius and $\mathfrak{F}$ is high. This is again not expected for poor models of genomes.

In practice, we estimate $\mathfrak{F}$ by calculating the ratio of median RAI score and the RAI score of $\lceil 0.1n \rceil^{th}$ fragment (where n is the number of fragments in the class) in descending order in a class of fragments assigned to a model. The underlying reason for this estimation is the ratios of sample numbers estimate the integrals in a Monte Carlo sense.

Simulated Annealing Strategy

At the end of each LBG epoch, we detect the weak models and disturb them so that we can rerun the procedure as follows: If any two close models are both detected to be weak, these two clusters are merged since it is possible that they are two different models trying to describe the same genome. On the other hand, the weak model with the least total RAI score is split into two models since it might be a single model trying to describe multiple genomes. Iteratively following this procedure allows us to correct the weak models and improve the optimization.

```
PROCEDURE::SA based on model fittness

  1. set the objective function to -$\infty$
  2. run LBG
  3. detect weak models using model fitness criteria
  4. if the change in the objective is not significant or there
     are no weak models, TERMINATE
  5. merge the closest two weak models; if there is one, randomly
     disturb it
  6. split the weak model with the smallest RAI score into two
  7. GOTO 2
```

We have empirically determined the model fitness threshold as 4.1 calculating over good ($< R_{0.2}$ away from actual centroid) and poor ($> R_{0.2}$ away from actual centroid) models. The simulated annealing stops when all the models are determined to be good.

7.4 Results

Unsupervised RAIphy performs as an unsupervised binning program, since it takes metagenome sequences as input and returns taxonomic clusters. In order to evaluate its performance, we compared unsupervised RAIphy with existing binning programs. To simulate metagenomes, we have picked a set of random genomes from NCBI RefSeq database with random relative abundance values. Minimum coverage of 3X with 400 bp average read length is considered to simulate the current pyrosequencing technology. We used a varying number of genomes for each experiment to represent metagenomes of different complexity. We repeated each experiment 100 times to obtain average accuracy values.

The quality of clustering depends on several factors. We adopted two measures of quality to assess the accuracy of clustering. The first measure is recall, which assumes that if a cluster has the most elements of a class, then it represents the class. The true positive calculation according to this assignment

$$Recall = \frac{1}{N}\sum_{j}\max_{i} C_{ij} \tag{7.8}$$

where C_{ij} is the number of sequencing reads in cluster i belonging to taxon j, and N is the total number of reads in the metagenome.

Similarly, precision assumes a cluster has the label of the most dominantly populated class in it, and measures the true positives with this assumption:

$$Precision = \frac{1}{N}\sum_{i}\max_{j} C_{ij} \tag{7.9}$$

The clustering accuracies are compared with the results of three well-known unsupervised binning algorithms: LikelyBin, CompostBin, and SCIMM. The average recall and precision values show for metagenomes of varying abundance show that unsupervised RAIphy improves the binning accuracy significantly. We have seen that for metagenomes simulations consisting of 5, 20 and 50 species, unsupervised RAIphy binning stayed above 80% recall performance on average. Similarly, the precision value was never observed to be below 75% for the entire set of experiments. The average performance with these experiments are shown in the graphs 7.7, 7.8, and 7.9.

It is known that new generation sequencing technologies have relatively higher

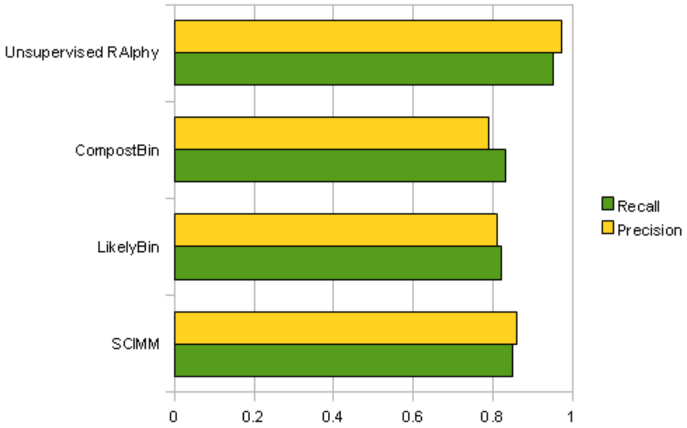

Figure 7.7: The average binning performance of unsupervised RAIphy algorithm compared with LikelyBin, CompostBin, and SCIMM programs for metagenome sets of 5 organisms. 100 experiments with varying population abundance is performed.

sequencing error rates than Sanger sequencing [215]. To see the effect of sequencing errors on our unsupervised binning method, we simulated sequencing with changing error rates. An experiment on 5 genome dataset with 400 bp reads have shown that unsupervised RAIphy is very robust for sequencing errors. This can be considered as an advantage provided by genome signatures for averaging out the sequence errors and smoothing their effects.

error rate %	Recall	Precision
0	0.951	0.969
1	0.947	0.952
3	0.928	0.944
5	0.916	0.941

Table 7.2: The performance drop in metagenome binning of 5 genomes with sequencing error. The unsupervised RAIphy algorithm exhibits a robust nature to sequencing errors. The recall and precision values do not fall under 0.9 even with 5% error rate.

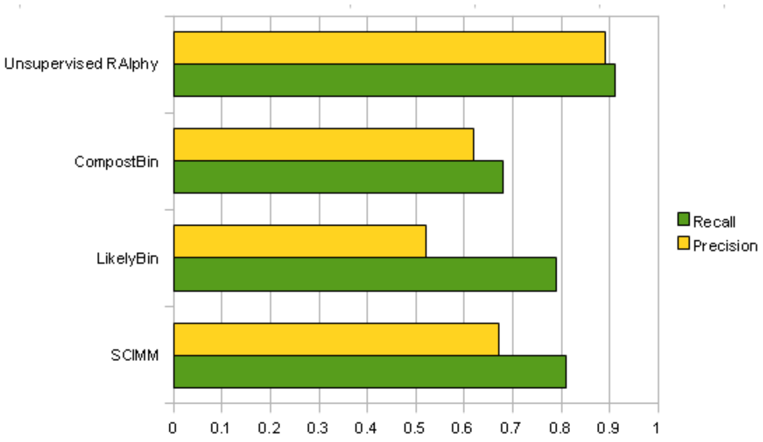

Figure 7.8: The average binning performance of unsupervised RAIphy algorithm compared with LikelyBin, CompostBin, and SCIMM programs for metagenome sets of 20 organisms. 100 experiments with varying population abundance is performed.

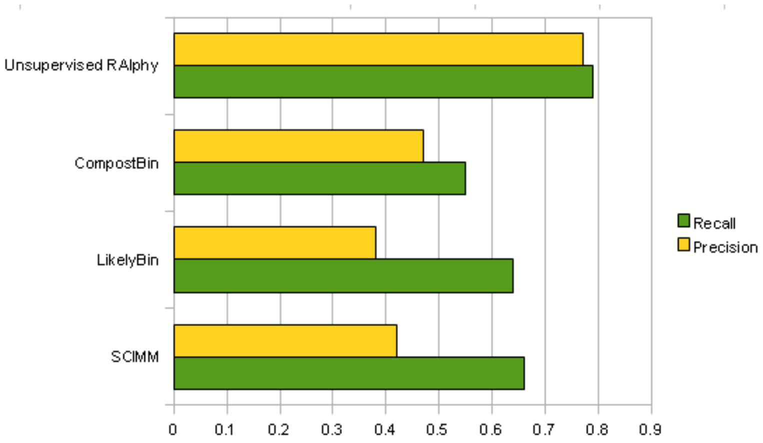

Figure 7.9: The average binning performance of unsupervised RAIphy algorithm compared with LikelyBin, CompostBin, and SCIMM programs for metagenome sets of 50 organisms. 100 experiments with varying population abundance is performed.

Chapter 8

Summary and Future Work

This work concentrates on a mathematical characterization of DNA sequences called genome signatures. What distinguishes genome signatures from other types of mathematical characterizations is their species specific and pervasive nature. The premise of genome signatures is that they result in unique representations for each species; this is associated with the species specific feature. Moreover, the associated representation can be deduced from a random genomic fragment of the corresponding species; this is associated with the pervasive feature. The genome signature concept provides various opportunities for computational biology applications in diverse fields such as metagenomics, phylogenetics, and evolutionary biology. For example a genomic fragment of an unknown source can be assigned to its origin of species using genome signature similarities between the fragment and the known species.

It was shown that well-known DNA sequence characterizations such as GC content, amino acid usage and synonymous codon usage can be considered as direct or approximate functions of oligonucleotide occurrence frequencies of a DNA sequence. We have introduced some possible functions of oligonucleotide frequencies as a class of genome signatures which emphasize the short term dependencies of the nucleotides forming the double helix. They are distinguished from the signatures which characterize longer term correlations or some other complexity features of genomes. Although most of them are in the same class, different genome signatures emphasize different conserved properties of genomic sequences.

An important observation on genome signatures is that the distance metrics measuring the similarity of DNA sequences significantly affect the pervasiveness and species specificity. For example, the ANOVA tests performed on different genome signatures show that simple signatures (e.g. oligonucleotide frequencies) can be more specific than signatures featuring more complicated attributes of genomes (e.g. abundance indices, which measure the deviation from randomness) in a Euclidean distance sense. In fact, the signatures which appear to be weak for one distance metric can contain significant information about its genome. It is possible to exploit the information contained in a characterization more efficiently with a better choice of similarity measurement.

We have seen that signatures measuring the relative abundance of oligonucleotides appear to be less species specific when Euclidean distance is considered. Perhaps, because of this reason, most computational biology applications prefer oligonucleotide content over functions of oligomer counts. We studied probabilistic measurements rather than absolute metrics, and we observed that the information contained in oligonucleotide relative abundance measurements can be utilized more efficiently. As a result, better pervasiveness and specificity is obtained. The novel signature is named as Relative Abundance Index (RAI).

The improvement promised by RAI has been validated on a metagenome analysis application: the RAIphy program. With the introduction of a semi-supervised nature using an iterative refinement algorithm, RAIphy provides accurate metagenome binning for a large range of DNA fragment lengths. We reported that RAIphy performs better than current compositional binning methods for the sequencing read length of contemporary sequencing technology (100 bp-1000 bp), and it is competitive with similarity-search based methods for the same range. For longer DNA contigs, RAIphy outperformed the existing binning programs. RAIphy's compact, fast, and parallelizable nature makes it suitable as a taxonomy assignment module of metagenome analysis pipelines.

As an alternative application using genome signatures with a probabilistic measurement approach, we have designed an unsupervised binning method. This method, called unsupervised RAIphy, uses Markov models with model fitness measurements and RAI measurements to taxonomically cluster sequence reads. The framework developed for the unsupervised binning method replaces the sequential paradigm of metagenome analysis. Instead of assembling genomes first and binning subsequently, we employed a convolved process of pseudoassembly and clustering. Genome signatures applications involve in both phases for pseudocontig generation and taxonomic clustering. The results of unsupervised RAIphy offer significant improvement over current unsupervised binning techniques.

Use of genome signatures with carefully defined distance metrics promises a great potential for many areas. Especially we have observed that the pervasiveness and specificity of signals embedded in genomic sequences might be potentially greater than the expectations. The related information contained in biological sequences can be exploited more carefully with further study. Some potentially novel applications of genome signatures could be listed as follows.

Metagenome assembly

The results obtained by the novel unsupervised binning approach is promising for the development of parallelized algorithms in order to assemble metagenomics data from different genomes simultaneously. A straightforward procedure can be cascading whole genome assemblers to the binning pipeline (e.g. unsupervised RAIphy) in order to produce genome assemblies.

The unsupervised RAIphy algorithm deploys concepts of genome assembly to be used in metagenome binning. A more intimate communication between assembly and binning might be useful for a monolithic approach instead of a pipeline. In this context, a monolithic method means a parallelizable recursive algorithm for genome assembly. This includes the assembly of small contigs and clustering them repeatedly until no further contig assembly is achievable. A recursive assembly in this manner was proposed for single genome assembly [216], and improvement in accuracy was observed for even single genome assembly.

Here, also the species specificity of compositional features contributes to dividing the problem into subproblems in different genomes. In this manner, an alternative strategy could be employing contig assembly on pseudocontig groups and clustering them with unsupervised RAIphy. Repeating this procedure iteratively similar to [216] until no more assembly is possible, will form a recursive strategy for metagenome assembly.

The accuracy of metagenome binning and assembly methods is expected to increase with DNA read length since longer DNA fragments contain more information. The fragment walk procedure results from unsupervised RAIphy suggests that greedy contig generation accuracies with 10 Kbp and 50 Kbp fragments is observed to be high. In fact, for this range of DNA read lengths, the binning process is highly accurate. Moreover the genome assembly is also much easier. Although we have designed our research methodology for second generation high-throughput sequencing which provide a large number of short DNA reads, the accuracy increases further with longer fragments. Prospective third generation genome sequencing which might be available soon [217-219] will be capable of yielding inexpensive DNA reads in the range of 10 Kbp-50 Kbp. Sufficient deep sequencing of microbiomes by the prospective sequencing technology and the analysis of data using the proposed approach has the potential to open the door for *ab initio* metagenome annotation in which the biodiversity and genomes existing in the environment are explored *in silico* from sequencing data.

Computational comparative metagenomics

The biological diversity studies associated with microbial communities have revealed that the relationship of microbiata with host organisms or an abiotic environment is related with the composition of the communities. The human microbiome project [220] enabled the focusing on whether and how the microbial communities effect human health. Investigation of existence/absence of bacteria and its contribution to human disease [221], detecting the microbial elements of human obesity [222], abundance differences between human infant and mature gut microbiomes reflecting the difference in digestion patterns [223, 224], the effects of mammalian microbiomes on the host cholesterol metabolism [225] are some examples emphasizing the importance of microbiomes to human health.

The characterization of microbial diversity has been performed with certain methods identifying and categotizing microbial organisms taxonomically. The most popular characterization technique has been ribotyping using 16S RNA genes, Multilocus Sequence Typing, and the use of marker genes. Recently developed methods MEGAN, CARMA, SONs [226], Libshuff [227], and Metastats [228] have limitations in metagenome characterization. These limitations are mainly due to low resolution of marker gene approaches and being confined to currently explored taxa.

In order to characterize microbial communities more carefully with exploiting metagenomes and gathering more information, characterizing microbiomes by *metagenome signatures* and conduct comparative metagenomics using this mathematical characterization is a potential approach to microbiome studies. While new generation DNA sequencing technologies provide feasibility of deep sequencing of metagenomes, inferring compositional maps of metagenomes is now achievable.

Using metagenome signatures in order to compare microbial communities will provide several advantages. First of all, the data we gather to process constitute a compositional image of the metagenome instead of taxonomic composition information. We can make use of it by discretizing the map with a desired resolution and digitally process it in the metric space where the signatures are defined. As the metagenome signatures, models derived from a pseudocontig generation process can be used.

Since every taxon has an index (or similarly ID) and an abundance value as a result of digital signal processing of metagenome signatures, methods from gene expression research can be exported. The analogy can be conducted as: the genes are replaced by genomes (or taxonomic units of some resolution), the expression values are replaced by relative abundance in the population and the hypothesis "gene expression patterns contain information about the state of cells" is replaced by the hypothesis “the composition patterns of microbiomes contain information about the state of organisms”. It is possible to employ serial analysis for detecting which microbes are changing the composition similar to the serial analysis of gene expression [229]. Moreover, multivariate statistics such as vector clustering (a vector consists of multiple taxa) can be employed to observe the positive or negative correlations for symbiosis estimation studies. There are several feature selection techniques to detect the active variables in a process. Those feature selection methods can be used for detecting active components of microbiome OTUs effective in a process. A number of machine learning procedures have been useful for the detection of pathology (such as cancer) using pattern recognition in clinical gene expression data [230]. Adopting the concepts from that know-how, clinical samples can be used for training classes and the detection of pathology or hypothesis testing involving evolutionary characteristics of microbiomes.

Capturing genome signature data in time series as a function of molecular evolution

The idea of capturing the compositional characteristics of a microbial community using genome signatures is representing it with a set of vectors in a metric space which is phylogenetically meaningful. Considering the hypothesis that genome signatures are driven by evolutionary processes, metagenome signature data are snapshots of a phase portrait at a given time. According to this hypothesis, genome signatures migrate/diffuse in the metric space during the course of evolution. Attempts to model these evolutionary dynamics using the mentioned phase-portrait approach is a significant both for i) Estimating the ecological dynamics, and ii) as a mathematical approach for modeling microbial evolution and coevolution of communities. An approach for such an attempt is using computer vision methods such as motion analysis and diffusion dynamics [231]. A series of clinical samples form images of time series and models derived from these data will be utilized to see whether evolution can be predictable based on model fitness as well as estimating patterns and trends in evolution under certain treatment.

Evolutionary implications of genome signatures

Genome signatures capture the compositional features of DNA content in organisms. They reflect the total net response of a genome to its environment. The relationship of environment and genome composition has been investigated for compositional features such as GC content, amino acid usage, synonymous codon usage and genome signatures. In this work, we explored that better modeling is achievable by defining signatures emphasizing various compositional features with careful selection of distance metrics. Following this line of thought, better correlations with environmental factors and genome composition can be addressed. For instance, support vector regression of abundance index profiles is a candidate for a environmental factor-genome composition investigation. Better correlations between environmental factors (e.g. optimal growth temperature, habitat, respiratory behavior, nutrition, etc.) and composition or discovery of unknown relations might be valuable for better understanding of organism-environment relationships, as well as molecular evolution.

It has been discovered that sequenced genomes of organisms from all domains of life have specific k-distributions of oligonucleotides. That kind of distribution can be modeled by double-Pareto-lognormal distributions [232], meaning that different Pareto distributions are fitted for both tails and a lognormal distribution is observed in the middle section. Double-Pareto-lognormal distribution successfully fits to the oligonucleotide occurrence of different organisms, from prokaryotes to higher eukaryotes with different parameters [233]. This distribution is associated with a random evolution by duplication model. According to that model, the evolution initiates with a short random genomic sequence. This sequence can be an outcome of Bernoulli process in which no correlation exists between the nearby nucleotides. Then the genome starts to expand with random copy-paste editions. A random section of the genome is copied and inserted to another random location with random point mutations. The simulations of this simple process is employed with random seed of 1000bp sequences and duplication of 25-33 bp sections until a genome size is reached [232,233]. The k-distribution statistics of the simulated genomes match surprisingly well with the corresponding real-life genomes. Therefore a neutralist evolution model is proposed with random duplications of genomic segments in the early age of evolution to a last universally known common ancestor. The optimal initial seed length, which is around 1000 bp agrees with the "RNA-world hypothesis for the origin of life" [233,234]. Therefore, a from a small stable RNA sequence, the small sections around 25-31 bp are copied by ribzymes and growth of genome followed that strategy.

The duplicative evolution hypothesis is supported by the experiments simulating the process which yield similar k-distributions. However, according to this argument, the statistics of real genomes and the genomes generated by duplicative evolution simulations match for short term correlations. A computational hypothesis testing procedure considering longer term correlations can be considered by employing the corresponding genome signatures. Average mutual information profiles and correlation strength signatures, which estimate the longer term correlations of DNA sequences can provide broader insights to the RNA-world hypothesis for the origin of life.

Bibliography

1. O. U. Nalbantoglu, S. F. Way, S. H. Hinrichs, K. Sayood. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics.* Jan 31;12:41, 2011.

2. C. H. Cannon, C .Kua, E. K. Lobenhofer, P. Hurban. Capturing genomic signatures of DNA sequence variation using a standard anonymous microarray platform. *Nucleic Acids Research*, 34. Art. No. e121, 2006.

3. K. J. Livak, S. J. Flood, J. Marmaro, W. Giusti, K. Deetz. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* 4: 357362, 1995.

4. A. M. Phillippy, J. A. Mason, K. Ayanbule, D. D. Sommer, E. Taviani, A. Huq, R. R. Colwell, I. T. Knight, S. L. Salzberg. Comprehensive DNA signature discovery and validation. *PLoS Comput Biol*, 3(5):e98, 2007.

5. P. D. Hebert, M. Y. Stoeckle, T. S. Zemlak and C .M. Francis. Identification of birds through DNA barcodes. *PLoS Biol*, 2:e312, 2004.

6. R. D. Ward, T. S. Zemlak, B. H. Innes, P. R. Last and P. D. Hebert. DNA barcoding Australias fish species. *Philos Trans R Soc Lond B Biol Sci*, 360:18471857, 2005.

7. M. Vences, M. Thomas, A. van der Meijden, Y. Chiari, D. R. Vieites. Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology 2*: article 5, 2005.

8. S. E. Miller. DNA barcoding and the renaissance of taxonomy. *Proc Natl Acad Sci USA*, 104:47754776, 2007.

9. K. H. Chu, C. P. Li and J. Qi. Ribosomal RNA as molecular barcodes: A simple correlation analysis without sequence alignment. *Bioinformatics*, 22:16901701, 2006.

10. K. H. Chu, M. Xu, C. P. Li. Rapid DNA barcoding analysis of large datasets using the composition vector method. *BMC Bioinformatics*, 10(Suppl 14):S8, 2009.

11. A. Nakabachi, A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, et al. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* 314: 267, 2006.

12. K.Y. Lee, R. Wahl, E. Barbu. Contenu en bases puriques et pyrimidiques des acides desoxyribonucleiques des bacteries. *Ann. Inst. Pasteur.* 91, 212224, 1956.

13. H. Ochman, and J. G. Lawrence. Phylogenetics and the amelioration of bacterial genomes, p. 26272637. In F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), Escherichia coli and Salmonella: cellular and molecular biology, 2nd ed., vol. 2. ASM Press, Washington, D.C, 1996.

14. W. Saenger. Principles of Nucleic Acid Structure. Springer-Verlag, New York, 1984.

15. E. Freese. On the evolution of base composition of DNA. *J. Theor. Biol.*, 3, 82101, 1962.

16. N. Sueoka. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA.*, 48, 582 592, 1962.

17. T. Y. Cheng, N. Sueoka. Heterogeneity of DNA in density and base composition. *Science*,Sep 20;141:1194-6, 1963.

18. R. Cavicchioli. Cold-adapted archaea. *Nature Reviews Microbiology*, 4:331-343, 2006.

19. D. A. Hickey, G. A. C. Singer. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol.*, 5(10):117, 2004.

20. C. McEwan, D. Gatherer and N. McEwan. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* 128, 173178, 1998.

21. J. G. Bragg, C. L. Hyder. Nitrogen versus carbon use in prokaryotic genomes and proteomes. *Proc Biol Sci.*, 271(Suppl 5):S374S377, 2004.

22. J. G. Bragg et al. Variation among species in proteomic sulphur content is related to environmental conditions. *Proc. Biol. Sci.*, 273, 12931300, 2006.

23. B. S. Berlett, E. R. Stadtman. Protein Oxidation in Aging, Disease, and Oxidative Stress. *The Journal of Biological Chemistry*, 272:2031320316, 1997.

24. R. Sandberg,C. I. Brnden,I. Ernberg,J. Cster. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. *Gene*, Jun 5;311:35-42, 2003.

25. R. Grantham. Workings of the genetic code.*Trends Biochem.Sci.*5:327-331, 1980.

26. R. Grantham,C. Gautier,M. Gouy,R. Mercier,A. Pav. Codon catalog usage and the genome hypothesis.*Nucleic Acids Res.*,Jan 11;8(1):r49r62, 1980.

27. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, R. Mercier. Codon catalog usage is a genome strategy modulated for gene expressivity.*Nucleic Acids Res.*, Jan 10;9(1):r43r74, 1981.

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

29. A. Carbone, A. Zinovyev, F. Kepe‘s. Codon Adaptation Index as a measure of dominating codon bias. *Bioinformatics* 19:20052015, 2003.

30. G. D'Onofrio, D. Mouchiroud, B. Assani, C. Gautier, G. Bernardi. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *Journal of molecular evolution*,32(6):504-10, 1991.

31. J. R. Lobry. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, 205:309316, 1997.

32. P. G. Foster, L. S. Jermiin, D. A. Hickey. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol.*, Mar;44(3):282-8, 1997. 1991.

33. N. Sueoka. Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harb. Symp. Quant. Biol.* 26:3543, 1961.

34. A. A. Adzhubei, I. A. Adzhubei, I. A. Krasheninnikov, S. Neidle. Nonrandom usage of degenerate codons is related to protein three-dimensional structure. *FEBS Lett*, 399:78-82, 1996.

35. T. Xie, D. Ding, X. Tao, D. Dafu. The relationship between synonymous codon usage and protein structure [published erratum appears in FEBS Lett, Oct 16;437(1-2):164]. *FEBS Lett* 1998, 434:93-96, 1998.

36. S. K. Gupta, S. Majumdar, T. K. Bhattacharya, T. C. Ghosh. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem Biophys Res Commun*, 269:692-696, 2000.

37. S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, H. H. McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci.* 101: 3480-3485, 2004.

38. R. Knight, S. Freeland, L. Landweber. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, 2001.

39. G. A. Palidwor , T. J. Perkins, X. Xia. A general model of codon bias due to GC mutational bias. *PLoS One*, Oct 27;5(10):e13431, 2010.

40. M. Zama. Codon usage and secondary structure of mRNA. *Nucleic Acids Symp Ser*, 22:93-94, 1990.

41. A. Carbone, F. Kepes, A. Zinovyev. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Mol Biol Evol.*, 22:547561, 2005.

42. J. Josse, A. D. Kaiser, and A. Kornberg. Enzymatic synthesis of deoxyribonucleic acid: VIII. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *The Journal of Biological Chemistry*, 236:864-875, 1961.

43. M. N. Swartz, T. A. Trautner, and A. Kornberg. Enzymatic synthesis of deoxyribonucleic acid: XI. further studies on nearest neighbor base sequences in deoxyribonucleic acid. *The Journal of Biological Chemistry*, 237:1961-1967, 1962.

44. J. H. Subak-Sharpe. Base doublet frequency patterns in the nucleic acid and evolution of viruses. *British Medical Bulletin*, 23:161-168, 1967.

45. G. J. Russell and J. H. Subak-Sharpe. Similarity of the general designs of protochordates and invertebrates. *Nature*, 266:533-536, 1977.

46. J. M. Morrison, H. M. Keir, J. H. Subak-Sharpe, and L. V. Crawford. Nearest neighbour base sequence analysis of the deoxyribonucleic acids of a further three mammalian viruses: Simian virus 20, human papilloma virus and adenovirus type 2. *Journal of General Virology*, 1:101-108, 1967.

47. C. Burge, A. M. Campbell, and S. Karlin. Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences*, 89:1358-1362, February 1992.

48. S. Karlin and L. R. Cardon. Computational DNA Sequence Analysis. *Annual Review of Microbiology*, 48:619-654, 1994.

49. S. Karlin and I. Ladunga. Comparison of eukaryotic genomic sequences.

Proceedings of the National Academy of Sciences, USA, 91:12832-12836, December 1994.

50. S. Karlin, E. S. Mocarski, and G. A. Schachtel. Molecular Evolution of Herpesviruses - Genomic and Protein Sequence Comparison. *Journal of Virology*, 68:1886-1902, March 1994.

51. S. Karlin, I. Ladunga, and B. E. Blaisdell. Heterogeneity of genomes: Measure and values. *Proceedings of the National Academy of Sciences, USA*, 91:12387-12841, 1994.

52. S. Karlin, J. Mrazek, and A. M. Campbell. Compositional Biases of Bacterial Genomes and Evolutionary Implications. *Journal of Bacteriology*, 179:3899-3913, June 1997.

53. S. Karlin, A. M. Campbell, and J. Mrazek. Comparitive DNA Analysis across Diverse Genomes. *Annual Review of Genetics*, 32:185-225, December 1998.

54. S. Karlin and C. Burge. Dinucleotide relative abundance extremes: A genomic signature. *Trends in Genetics*, 11:283-290, July 1995.

55. S. Karlin and J. Mrazek. Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 1022710232, September 1997

56. F. L. Bai, Y. Z. Liu. Wang TM. A representation of DNA primary sequences by random walk. *Math Biosci*, Sep;209(1):282-91, 2007.

57. M. A. Gates. Simple DNA sequence representations. *Nature*, 316, 219, 1985.

58. M. Leong, S. Morgenthalar. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.*, 21, 503, 1995.

59. M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M. R. Dudek, S. Cebrat. High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. *BMC evolutionary biology*, 17:1-13, . 2001.

60. H.J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18:2163-2170, 1990.

61. R. L. Devaney, An Introduction to Chaotic Dynamical Systems. Addison Wesley, Redwood City, California, 1989.

62. P. J. Deschevanne, A. Giron, J. Vilain, A. Vaury, and B. Fertil. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution*, 16:1391-1399, 1999.

63. P. J. Deschavanne, A. Giron, J. Vilain, C. Dufraigne, B. Fertil. Genomic signature is preserved in short DNA fragments. BIBE 2000 IEEE International Symposium on bio-informatics & biomedical engineering. Washington, D.C., USA, 2000.

64. Y. Wang, K. Hill, S. Singh, and L. Kari. The spectrum of genomic signatures: From dinucleotides to chaos game representation. *Gene*, 346:173-185, 2005.

65. R. W. Jernigan, R. H. Baran. Pervasive properties of the genomic signature. *BMC Genomics*, 3, 23, 2002.

66. C. K. Peng , S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley. Long-range correlations in nucleotide sequences. *Nature*, Mar 12;356(6365):168-70, 1992.

67. W. Li, K. Kaneko. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* 17, 655660, 1992.

68. R. F. Voss. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* 68, 38053808, 1992.

69. S. Karlin, V. Brendel. Patchiness and correlations in DNA sequences. *Science* 259, 677680, 1993.

70. S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, H. E. Stanley. Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev.* E 51, 50845091, 1995.

71. D. Holste , I. Grosse, S. Beirer, P. Schieg, H. Herzel. Repeats and correlations in human DNA sequences. *Phys. Rev.* E 67, 061913, 2003.

72. H. Herzel , W. Ebeling, A. Schmitt. Entropies of biosequences: the role of repeats. *Phys. Rev.*, E 50, 5061 5071, 1994.

73. D. Holste, I. Grosse, S. Buldyrev, H. Stanley, H. Herzel. Optimization of coding potentials using positional dependence of nucleotide frequencies. *J. Theor. Biol.* 206, 525537, 2000.

74. I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver, H.

Stanley. Analysis of symbolic sequences using the Jensen Shannon divergence. *Phys. Rev.*, E 65, 041905, 2002.

75. M. Berryman, A. Allison, D. Abbott. Mutual information for examining correlations in DNA. *Fluctuation Noise Letters* 4, 237246, 2004.

76. P. Jacobs, P. Lewis. Discrete time series generated by mixtures III: autoregressive processes (DAR(p)). Tech. Rep. NPS55-78-022, Naval Postgraduate School, Monterey, California, 1978.

77. P. Jacobs, P. Lewis. Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Ser. Anal.*, 4, 1936, 1983.

78. M. Dehnert, W.E. Helm, M. T. Hqtt. A discrete autoregressive process as a model for short-range correlations in DNA sequences. *Physica A* 327, 535553, 2003.

79. M. Dehnert, W. Helm, M.T. Hutt. Information theory reveals large scale synchronisation of statistical correlations in eukaryote genomes. *Gene*, 345:81-90, 2005.

80. M. Dehnert, W. Helm, M. T. Hutt. Informational structure of two closely related eukaryote genomes. *Physical Review E*, 74:021913-1-021913-9, 2006

81. C. Shannon. A Mathematical Theory of Communication. *Bell Syst Tech J*, 27:379-423. 62365, 1948.

82. B. Korber, R. Farber, D. Wolpert, A. Lapedes. Covariation of Mutations in the V3 Loop of Human Immunodeficiency Virus Type I Envelope Protein: An Information Theoretic Analysis. *Proc Natl Acad Sci*, 90:7176-7180, 1993.

83. R. Roman-Roldan, P. Bernaolo-Galvan, J. Oliver. Application of Information Theory to DNA Sequence Analysis: A Review. *Pattern Recognition*, 29(7):1187-1194, 1996.

84. B. Giraud, A. Lapedes, L. Liu. Analysis of Correlations Between Sites in Models of Protein Sequences. *Phys Rev E*, 58(5):6312-6322, 1998.

85. H. Herzel, I. Grosse. Correlations in DNA Sequences: The Role of Protein Coding Segments. *Phys Rev E*, 55:800-810, 1997.

86. I. Hofacker, M. Fekete, P. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, 319:1059-1066, 2002.

87. S. Lindgreen, P. Gardner, A. Krogh. Meauring covariation in RNA alignments: physical realism improves information measure. *Bioinformatics*, 22:2988-2995, 2006.

88. M. Bauer, S. Schuster, K. Sayood. The average mutual information profile as a genomic signature. *BMC Bioinf.*, 9, 2008.

89. H. Herzel, O. Weiss, E. N. Trifonov. 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*, Mar;15(3):187-93, 1999.

90. E. N. Trifonov, and J. L. Sussman. The pitch of chromatin DNA is reflected in its nucleotide sequence. *PNAS* 77, 38163820, 1980.

91. M. Berryman, A. Allison, D. Abbott. Mutual Information for Examining Correlations in DNA. *Fluctuation and Noise Letters*, 4(2):L237-L246, 2004.

92. M. Dehnert, R. Plaumann, W. E. Helm, M. T. Htt. Genome phylogeny based on short-range correlations in DNA sequences. *J Comput Biol.* Jun;12(5):545-53, 2005.

93. S. L. Salzberg, et al. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, 26:544548, 1998.

94. A. L. Delcher, K. A. Bratke, E. C. Powers, S. L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23, 673679, 2007.

95. D. Dalevi, D. Dubhashi, and M. Hermansson. Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. *Bioinformatics*, 22:517-522, March 2006.

96. I. Saeed, S. K. Halgamuge. The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. *BMC Genomics.* Dec 3;10 Suppl 3:S10, 2009.

97. N. R. Pace. A molecular view of microbial diversity and the biosphere. *Science* 276: 734740, 1997.

98. M. S. Rappe , S. J. Giovannoni. The uncultured microbial majority. *Annu Rev Microbiol* 57: 369394, 2003.

99. R.D. Fleischmann , M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science* 269: 496512, 1995.

100. National Research Council of the National Academies. The dawning of a new microbial age. in The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet p. 2 (The National Academies Press, Washington, DC, 2007).

101. M. Achtman, and M. Wagner. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.*, 6:431-440, 2008.

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

103. E. DeLong. Microbial Community Genomics in the Ocean. *Nature Reviews*, 3:459469, 2005.

104. E. DeLong, C. Preston, T. Mincer, V. Rich, S. Hallam, N. U. Frigaard, A. Martinez, M. Sullivan, R. Edwards, B. Brito, S. Chisholm, D. Karl. Community Genomics Among Stratified Microbial Assemblages in the Oceans Interior. *Science*, 311:496503, 2006.

105. S. M. Barns, R. E. Fundyga, M. W. Jeffries, N. R. Pace. Remarkable archeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci USA*, 91:16091613, 1994.

106. R. Huber, H. Huber, K. O. Stetter. Towards ecology of hyperthermophiles: biotypes, new isolation strategies and novel metabolic properties. *FEMS Microbiol Rev*, 24:615623, 2002.

107. B. C. Christner, B. H. Kvitko, J. N. Reeve. Molecular identification of Bacteria and Eukraya inhabiting an Antarctic cryoconite hole. *Extremophiles*,

7:177183, 2003.

108. S. Bellnoch. Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environmental Microbiology*, 4:349360, 2002.

109. P. Lorenz, J. Eck. Metagenomics and industrial applications. *Nature Reviews*, 3:510516, 2005.

110. C. Schmeisser, H. Steele, W. Streit. Metagenomics, biotechnology with nonculturable microbes. *Appl Microbiol Biotechnol*, 75:955962, 2007.

111. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, J. I. Gordon. The Human Microbiome Project. *Nature*, 449:804810, 2007.

112. J. P. Noonan, G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Paabo, J. K. Pritchard, E. M. Rubin. Sequencing and Analysis of Neanderthal Genomic DNA. *Science*, 314 (5802):11131118, 2006.

113. R. E. Green, J. Krause, S. E. Ptak, A. W. Briggs and Ronan MTea: Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444:330336, 2006.

114. A. McHardy, I. Rigoutsos. Whats in the mix: phylogenetic classifcation of metagenome sequence samples. *Current Opinion in Microbiology*, 10:499503, 2007.

115. A. Andersson , M. Lindberg, H. Jakobsson, F. Backhed, P. Nyren ,L. Engstrand. Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing. *PLoS One*, 3(7):e2836, 2008.

116. S. Tringe, C. von Mering, A. Kobayashi, A. Salamov, K. Chen, H. Chang, M. Podar, J. Short, E. Mathur, J. Detter, P. Bork, P. Hugenholtz, E. Rubin.

Comparative Metagenomics of Microbial Communities. *Science*, 308:5547, 2005.

117. W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124-2129, 1999.

118. J. C. Wooley, A. Godzik, I. Friedberg. A primer on metagenomics. *PLoS Comput Biol* 26;6(2):e1000667, 2010.

119. D. Willner, M. Furlan, M. Haynes, R. Schmieder, F. E. Angly, et al. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 4: e7370, 2009.

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

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

122. M. Strous, E. Pelletier, S. Mangenot, T. Rattei, A. Lehner, M. W. Taylor, M. Horn, H. Daims, D. Bartol-Mavel, P. Wincker, V. Barbe, N. Fonknechten, D. Vallenet, B. Segurens, C. Schenowitz-Truong, C. Medigue, A. Collingro, B.

Snel, B. E. Dutilh, H. J. Op den Camp, C. van der Drift, I. Cirpus, K. T. van de Pas-Schoonen, H. R. Harhangi, L. van Niftrik, M. Schmid, J. Keltjens, J. van de Vossenberg, B. Kartal, H. Meier, D. Frishman, M. A. Huynen, H. W. Mewes, J. Weissenbach, M. S. Jetten, M. Wagner, D. Le Paslier. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, 440:790-794, 2006.

123. T. Woyke, H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E.M. Rubin, N. Dubilier. Symbiosis insights through metage- nomic analysis of a microbial consortium. *Nature* 443:950-955, 2006.

124. F. Warnecke, P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernan- dez, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz, J. R. Leadbetter. Metagenomic and func- tional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450:560-565, 2007.

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

126. F. Sanger, and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94:441-448, 1975.

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

128. Joint Genome Institute. (http://www.jgi.doe.gov/CSP/index.html).

129. J. C. Wooley, A. Godzik, I. Friedberg. A primer on metagenomics. *PLoS Comput Biol* 26;6(2):e1000667, 2010.

130. E. K. Wommack, J. Bhavsar, J. Ravel. Metagenomics: Read length matters. *Appl Environ Microbiol.*74(5):1453-1463, 2008.

131. E. S. Lander, M. S. Waterman Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2: 231-239, 1988.

132. E. R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.*, 9:387-402, 2008.

133. M. L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet.*, 11(1):31-46, 2010.

134. S. Fox, S. Filichkin, T. C. Mockler. Applications of ultra-high-throughput sequencing. *Methods Mol Biol.*, 553:79-108, 2009.

135. http://www.454.com/products-solutions/system-features.asp.

136. http://www.illumina.com/downloads/SQGAIIxspecsheet20409LR.pdf

137. http://www.appliedbiosystems.com/ABHome/applicationstechnologies/SOLiDSystemSeq

138. http://www.helicosbio.com/Technology/TrueSingleMoleculeSequencing/tSMStradePerforr

139. T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, G. L. Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72:5069-5072, 2006.

140. R. J. Case , Y. Boucher, I. Dahllof, C. Holmstrom, F. W. Doolittle, et al. Use of 16s rRNA and rpob genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* 73: 278-288, 2007.

141. C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315:1126-1130, 2007.

142. E. Mahenthiralingam, A. Baldwin, P. Drevinek, E. Vanlaere, P. Vandamme, et al. Multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* 1: e17. doi:10.1371, 2006.

143. M. T. Suzuki, and S. J. Giovannoni. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.*, 62:625-630, 1996.

144. F. von Wintzingerode, U. B. Gobel, and E. Stackebrandt. Determination of microbial diversity in environmental samples: pitfalls of PCR based rRNA analysis. *FEMS Microbiol. Rev.* 21:213-229, 1997.

145. M. Wu, J. A. Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, 9(10):R151, 2008.

146. M. Stark, S. A. Berger, A. Stamatakis, C. von Mering. MLTreeMap - accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics*, 11:461, 2010.

147. S. L. Salzberg, J. A. Yorke. Beware of misassembled genomes. *Bioinformatics*, 21:4320-4321, 2005.

148. P. L. Johnson, M. Slatkin. Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res*, 16:1320-1327, 2006.

149. P. Green. Phrap (www.phrap.org).

150. X. Huang, A. Madan. CAP3: A DNA sequence assembly program. *GenomeRes*,9:868-77, 1999.

151. S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre et al. ARACHNE: a whole-genome shotgun assembler. *Genome Res* 12: 177-189, 2002.

152. D. B. Jaffe, J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J. P. Mesirov, M. C. Zody, and E. S. Lander. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* 13: 91-96, 2003.

153. S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. ming Chia, et al. Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. *Science* 297: 1301-1310, 2002.

154. E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, et al. A whole-genome assembly of drosophila. *Science* 287: 2196-2204, 2000.

155. P. A. Pevzner, H. Tang, M. S. Waterman An eulerian path approach to DNA fragment as- sembly. *Proc Natl Acad Sci U S A* 98: 9748-9753, 2001.

156. M. J. Chaisson, and P. A. Pevzner. Short read fragment assembly of bacterial genomes. *Genome Res.*, 18: 324-330, 2008.

157. M. Pop. Genome assembly reborn: recent computational challenges. *Bioinformatics*, 4: 354- 366, 2009.

158. R. L. Warren, G. G. Sutton, S. J. Jones, R. A. Holt. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23: 500-501, 2007.

159. W. R. Jeck, J. A. Reinhardt, D. A. Baltrus, M. T. Hickenbotham, V. Magrini, E. R. Mardis, J. L. Dangl, C. D. Jones. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 23: 2942-2944, 2007.

160. J. C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.* 17: 1697-1706, 2007.

161. D. R. Zerbino, and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18: 821-829, 2008.

162. L. Krause et al. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res*, 36, 2230-2239, 2008.

163. D. H. Huson, A. F. Auch, J. Qi, S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Res.*, 17, 377-386, 2007.

164. D. H. Huson, D. C. Richter, S. Mitra, A. F. Auch, S. C. Schuster. Methods for comparative metagenomics. *BMC Bioinformatics*, 10(Suppl 1):S12, 2009.

165. S. F. Altschul. Gapped BLAST and PSI-BLAST: a new generation of protein database search program. *Nucleic Acids Res*, 125:3389-3402, 1997.

166. E. A. Dinsdale. Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One*, 3:e1584, 2008.

167. H. M. Monzoorul, S. Tarini, K. Dinakar, S. M. Sharmila. SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25:1722-1730, 2009.

168. A. Brady, S. L. Salzberg. Phymm and Phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nat Methods*, 6: 673-676, 2009.

169. R. Sandberg, G. Winberg, C. I. Branden, A. Kaske, I. Ernberg, J. Coster. Capturing whole-genome characteristics in short sequences using a naive Bayesian Classifier. *Genome Research*, 11:1404-1409, 2001.

170. D. Dalevi, D. Dubashi, M. Hermansson. Bayesian Classifiers for detecting HGT using xed and variable order Markov models of genomic signatures. *Bioinformatics*, March 22:517-522, 2006.

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

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

173. T. Kohonen. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43: 5969, 1982.

174. T. Kohonen, E. Oja, O. Simula, A. Visa, J. Kangas. Engineering applications of the self-organizing map. *Proc. IEEE* 84: 13581384, 1996.

175. G. J. Dick et al. Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10:R85, 2009.

176. K. U. Foerstner, C. von Mering, S. D. Hooper, P. Bork. Environments shape the nucleotide composition of genomes. *EMBO Rep*, 6:1208-1213, 2005.

177. H. Willenbrock, C. Friis, A. S. Juncker, D. W. Ussery. An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biol*, 7:R114, 2006.

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

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

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

181. C. K. K. Chan, A. L. Hsu, S. L. Tang, S. K. Halgamuge. Using Growing Self-Organising Maps to Improve the Binning Process in Environmental Whole-

Genome Shotgun Sequencing. *Journal of Biomedicine and Biotechnology*, 513701:10, 2008.

182. C. Martin, N. N. Diaz, J. Ontrup, T. W. Nattkemper. Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. *Bioinformatics*, 24:1568-1574, 2008.

183. C. Chan, A. Hsu, S. Halgamuge, S. Tang. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, 9:215, 2008.

184. S. Chatterji, I. Yamazaki, Z. Bai, J. Eisen. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In Research in Computational Molecular Biology, 12th Annual International Conference, RECOMB 2008, Singapore, March 30 - April 2, 2008. *Proceedings, Lecture Notes in Computer Science Volume 4955. Springer*; 2008.

185. W. J. Kent. Blat-the blast-like alignment tool. *Genome Res* 12(4), 656664, 2002.

186. A. Kislyuk, S. Bhatnagar, J. Dushoff, J. Weitz. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, 10:316, 2009.

187. Y. W. Wu, Y. Ye. A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-Tuples. *In Research in Computational Molecular Biology, of Lecture Notes in Computer Science.* Volume 6044. Edited by Berger B. Springer Berlin/Heidelberg; 535-549, 2010.

188. I. Sharon, A. Pati, V. M. Markowitz, et al. A statistical framework for the

functional analysis of metagenomes. In RECOMB 2009 . Springer Berlin / Heidelberg, Tucson, AZ,; 496-511, 2009.

189. H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, F. O. Glockner. Tetra: a webservice and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, 5:163, 2004.

190. B. Yang, Y. Peng, H. C. Leung, S. M. Yiu, J. C. Chen, F. Y. Chin. Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics.* Apr 16;11, 2010.

191. D. R. Kelley, S. L. Salzberg. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics.* Nov 2;11:544, 2010.

192. B. Schoelkopf and A. Smola. Learning with Kernels. MIT Press, Cambridge MA, 2002.

193. K. H. Chu, C. P. Li and J. Qi. Ribosomal RNA as molecular barcodes: A simple correlation analysis without sequence alignment. *Bioinformatics*, 22:16901701, 2006.

194. K. H. Chu, M. Xu, C. P. Li. Rapid DNA barcoding analysis of large datasets using the composition vector method. *BMC Bioinformatics* , 10(Suppl 14):S8, 2009.

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

196. J. Bohlin, E. Skjerve. Examination of genome homogeneity in prokaryotes using genomic signatures. *PLoS One.* 2009 Dec 2;4(12):e8113.

197. J. Bohlin, E. Skjerve, D. W. Ussery. Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol.* Apr 18;4(4):e1000057, 2008.

198. K. Sayood. Introduction to Data Compression, Third Edition. Morgan Kauffman- Academic Press, San Francisco, 2005.

199. C. Bishop. Pattern Recognition and Machine Learning. New York, NY: Springer; 2006.

200. Basu S, A. Banerjee, R. Mooney. Semi-supervised Clustering by Seeding. *In Proceedings of the Nineteenth International Conference on Machine Learning (ICML-2002)*, 2002.

201. K. D. Pruitt, T. Tatusova, D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35 Database: D61-D65, 2007.

202. G. E. Sims, S. R. Jun, G. A. Wu, S. H. Kim. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci US*, 106:2677-2682, 2008.

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

204. E. Bock, M. Wagner. In Oxidation of inorganic nitrogen compounds as an energy source. The Prokaryotes. Volume 3. 3 edition. New York, NY: Springer; 2006.

205. E. K. Wommack, J. Bhavsar, J. Ravel. Metagenomics: Read length matters. *Appl Environ Microbiol.*74(5):1453-1463, 2008.

206. K. Mavromatis, N. Ivanova, K. Barry, et al. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4:495-500, 2007.

207. J. C. Venter, K. Remington, J. F. Heidelberg, et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304:66-74., 2004.

208. E. W. Myers. Toward Simplifying and Accurately Formulating Fragment Assembly. *J. Comp. Bio*,2:275-90, 1995.

209. P. Medvedev, K. Georgiou, E. W. Myers et al. Computability and Equivalence of Models for Sequence Assembly. Workshop on Algorithms in Bioinformatics (WABI 2007). Philadelphia, PA: Springer, 2007

210. M. T. Tammi, E. Arner, E. Kindlund, and B. Andersson. Correcting errors in shotgun sequences. *Nucleic Acids Research*, 31:4663-4672, 2003.

211. S. J. Hallam, K. T. Konstantinidis, N. Putnam, C. Schleper, Y. Watanabe, J. Sugahara, C. Preston, J. de la Torre, P. M. Richardson, and E. F. DeLong. Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. *Proc. Natl. Acad. Sci. USA*, 103:18296-18301, 2006.

212. S. L. Simmons, G. Dibartolo, V. J. Denef, et. al. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biol*, 6:e177, 2008.

213. I. Grosse, P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, J. Oliver, H. E. Stanley. Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys. Rev. E Stat. Nonlin. *Soft Matter Phys.*, 65:041905, 2002.

214. Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantization design. *IEEE Transactions on Communications,* COM-28:84-95, Jan. 1980.

215. M. Ronaghi. Pyrosequencing sheds light onDNAsequencing. *Genome Res.*, 11, 311, 2001.

216. H. H. Otu and K. Sayood. A Divide and Conquer Approach to Sequence Assembly. *Bioinformatics*, 19(1):22-29, January 2003.

217. J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid, et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nano*, 4:265-270, 2009.

218. J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323: 133-138, 2008.

219. D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, et al. The potential and challenges of nanopore sequencing. *Nat Biotech*, 26: 1146-1153, 2008.

220. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, et al. The human microbiome project. *Nature* 449: 804810, 2007.

221. B. Rodriguez-Brito, F. Rohwer, R. A. Edwards. An application of statistics to comparative metagenomics. *BMC Bioinformatics* 7: 162, 2006.

222. R. E. Ley, P. J. Turnbaugh, S. Klein, J. I. Gordon. Microbial ecology: human gut microbes associated with obesity. *Nature* 444: 10221023, 2006.

223. K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 14: 169181, 2007.

224. P. A. Vaishampayan, J. V. Kuehl, J. L. Froula, J.L. Morgan, H. Ochman, and M.P. Francino. Comparative Metagenomics and Population Dynamics of the Gut Microbiota in Mother and Infant. *Genome Biol Evol.* 2010: 5366, 2010.

225. I. Martnez, G. Wallace, C. Zhang, R. Legge, A.K. Benson, T.P. Carr, E.N. Moriyama, and J. Walter. Diet- Induced Metabolic Improvements in a Hamster Model of Hypercholesterolemia Are Strongly Linked to Alterations of the Gut Microbiota. *Applied and Environmental Microbiology* 75: 4175-4184, 2009.

226. P. D. Schloss, J. Handelsman. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl Environ Microbiol* 72: 67736779, 2006.

227. D. R. Singleton, M. A. Furlong, S. L. Rathbun, W. B. Whitman. Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. *Appl Environ Microbiol* 67: 43744376, 2001.

228. J. R. White, N. Nagarajan, M. Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 5: e1000352, 2009.

229. V. E. Velculescu, L. Zhang, B. Vogelstein, K. W. Kinzler. Serial analysis of gene expression. *Science* 270: 484487, 1995.

230. S. B. Cho et al. Machine learning in DNA microarray analysis for cancer classification. *In Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, volume 19, 2003.

231. I. Cohen. Nonlinear variational method for optical flow computation. *In Proc. Eighth Scandinavian Conference on Image Analysis, Vol. 1, Troms, Norway*, pp. 523530, May 1993.

232. M. Csurs, L. No, G. Kucherov. Reconsidering the significance of genomic word frequencies. *Trends Genet.* Nov;23(11):543-6, 2007.

233. L. C. Hsieh, L. F. Luo and H. C. Lee. Evidence for Growth of Microbial Genomes by Short Segmental Duplications. *IEEE Proc. Comp. Sys. Bioinformatics*, 474-475, 2003.

234. W. Gilbert. The RNA world. *Nature*, p 618 v 319, 1986.