The Pennsylvania State University
The Graduate School

CLUSTERING ALGORITHMS FOR NEXT-GENERATION SEQUENCING DATA FROM HETEROGENOUS POPULATIONS

A Dissertation in
Computer Science and Engineering
by
Shruthi Prabhakara

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012

The dissertation of Shruthi Prabhakara was reviewed and approved* by the following:

Raj Acharya
Professor in Computer Science and Engineering
Dissertation Advisor, Chair of Committee

Webb Miller
Professor in Computer Science and Engineering
Professor of Biology

Padma Raghavan
Distinguished Professor in Computer Science and Engineering
Director of Institute of CyberScience

Mary Poss
Professor in Biology and in Veterinary and Biomedical Sciences

*Signatures are on file in the Graduate School.

Abstract

Next Generation Sequencing (NGS) technologies generate data more efficiently, economically and with a greater depth than ever before. NGS has opened up an array of possibilities for many applications including whole-genome sequencing, epigenetics, metagenomics and characterization of pathogens. Of these, the characterization of diversity of heterogeneous microbial environments such as metagenomes and viral populations has recently gained significant interest. Although a host of methods for whole genome assembly have been developed, reconstruction of heterogeneous populations using NGS data still remains a challenge. As compared to existing technologies, reads produced by NGS are typically shorter and more error-prone. The growth in the size of the datasets is fast outpacing the computational power needed to analyze it. Thus, many computational challenges arise while analyzing deep sequence data from heterogeneous populations[48]. The computational methods presented in this dissertation aim to analyze and quantify the genetic diversity within a heterogeneous population based on a set of deep sequencing reads.

A major challenge facing metagenomics is the development of tools for the characterization of functional and taxonomic content of vast amounts of short metagenome reads. Unlike single genome sequencing, assembly of a metagenome is intractable and is by large, an unsolved mystery. A crucial step in metagenomics that is not required in single genome assembly, is binning the reads belonging to a species. Clustering methods aim to identify the species present in the sample, classify the sequences by their species of origin and quantify the abundance of each of these species. The efficacy of clustering methods depends on the number of reads in the dataset, the read length and relative abundances of source genomes in the microbial community.

From clustering, to assembly, to annotation and function prediction, bioinformatics is posed with new challenges in handling noisy, huge and often partial

sequence data. Until now, methods to classify short sequences generated by NGS technologies have been relatively inaccurate. Most methods use sequence homology to assign reads to common ancestors. However, most extant databases are highly biased in their representation of true diversity, such methods fail to find homologs for reads derived from novel species.

To address the aforementioned problems, in the first part of dissertation, I have focused on methods to characterize and analyze the taxonomic content of vast amounts of short metagenome reads. The main contributions of the dissertation are: (i) A two-pass soft clustering semi-supervised method that is a hybrid of comparative and composition based methods. Such a hybrid approach is effective when evolutionarily close training genomes are available. (ii) An unsupervised Naive Bayes mixture model based on Poisson or Gaussian distributions to model reads within each bin. (iii) An unsupervised multivariate Bayesian mixture model based on Poisson and Multinomial distributions for clustering discrete sequence data. This method overcomes the bottleneck of above Naive Bayes by taking into account the conditional dependencies between the words within the reads. For the latter two methods, we present a two-way clustering approach to reduce the high-dimensionality and sparsity associated with the data. The method combines the words along the reads into word groups and constrains the parameters for words within the same group to be identical.

High genetic variability in viral populations plays an important role in disease progression, pathogenesis and drug resistance. The last few years has seen significant progress in the development of methods for reconstruction of viral populations using data from NGS technologies. These methods identify the differences between individual haplotypes by mapping the short reads to the reference genome. Much less has been published about resolving the population structure when an assembled reference genome is not well-defined, which severely limits the number of populations that can take advantage of these new technologies.

In the remainder of the dissertation, I have described a computational framework, called Mutant-Bin, for clustering individual haplotypes in a heterogeneous population and determining their prevalence, based on a set of deep sequencing reads. The main advantages of our method are that: (i) it enables determination of the population structure and haplotype frequencies when a reference genome is lacking; (ii) the method is unsupervised, the number of haplotypes does not have to be specified in advance; (iii) it identifies the polymorphic sites with derived nucleotides that co-occur in a subset of haplotypes and the frequency with which they appear in the viral population.

Table of Contents

List of Figures viii

List of Tables xi

Acknowledgments xii

Chapter 1
Introduction 1
1.1 Next Generation Sequencing 1
1.1.1 Genomics 2
1.1.2 Metagenomics 3
1.1.3 Heterogeneous Viral Populations 5
1.2 Research Contributions 6
1.2.1 Clustering Algorithms for Metagenomics 6
1.2.2 Haplotype Estimation of Viral Populations 10
1.3 Organization 10

Chapter 2
SIMCOMP: A Hybrid Clustering Algorithm 11
2.1 Related Work 12
2.2 Background and Motivation 14
2.3 Methods 16
2.3.1 Comparative Clustering 16
2.3.2 Composition Based Clustering 17
2.3.3 Definitions 18
2.3.4 SIMCOMP : Outline of the Algorithm 19
2.4 Results 20

2.4.1 Accuracy across Taxonomic Ranks . . . 21
2.4.2 Length of Oligomer . . . 23
2.4.3 Read Threshold . . . 25
2.5 Termite Metagenome . . . 25
2.6 Conclusion . . . 28

Chapter 3
A Naive Bayes Mixture Model **29**
3.1 Background and Motivation . . . 30
3.1.1 Multi-species Multi-dimensional Mixture of Distributions . 32
3.1.2 Parameter Estimation . . . 34
3.1.3 Word Grouping . . . 35
3.1.4 Naive Bayes Mixture of Multinomials . . . 37
3.2 Results . . . 38
3.2.1 Simulated metagenomes . . . 38
3.2.2 Real metagenome: Acid Mine Drainage Dataset . . . 43
3.3 Discussion . . . 44

Chapter 4
A Bayesian Mixture Model **51**
4.1 Background and Motivation . . . 51
4.2 Methods . . . 54
4.2.1 Bayesian Mixture of Poissons . . . 54
4.2.2 Two-Way Bayesian Mixture of Poissons . . . 56
4.2.3 Bayesian Mixture of Multinomials . . . 58
4.2.4 Two-Way Bayesian Mixture of Multinomials . . . 60
4.3 Results . . . 61
4.3.1 Datasets . . . 61
4.3.2 Accuracy Vs. Coverage . . . 63
4.3.3 Accuracy Vs. Number of clusters . . . 64
4.3.4 Accuracy Vs. Read length . . . 64
4.3.5 Accuracy Vs. Length of word . . . 64
4.4 Conclusion . . . 66

Chapter 5
Mutant Bin for Viral Haplotype Estimation **68**
5.1 Background and Motivation . . . 68
5.2 Methods . . . 71
5.2.1 Mixture of Poisson Distributions . . . 72

5.2.2 Cluster l-tuples using Variable Bandwidth Mean Shift Analysis . . . 74
5.2.3 Cluster l-tuples using Expectation Maximization . . . 76
5.2.4 Greedy Heuristic for Generating Set . . . 77
5.2.5 Inferring Phylogeny . . . 79
5.3 Experimental Results . . . 80
5.4 Conclusion . . . 89

Chapter 6
Conclusion **91**
6.1 Summary of Contributions . . . 92
6.2 Future Work . . . 94

Bibliography **104**

List of Figures

2.1 Organism level characterization of M1 dataset 21
2.2 Taxonomic Distribution Across Ranks (Phylum, Class, Order, Family 22
2.3 Average cluster purity across taxonomic ranks for ($RT_C = 15$ and $RT_F = 12$ and length of oligomer = 6, Number of Clusters = 2430) 23
2.4 Plot of percentage of non-singleton clusters for different values of purity with $RT_C = 25$ and $RT_F = 22$ and varying values of oligomers 24
2.5 Plot of percentage of non-singleton clusters for different values of purity with oligomer lenght = 6 and varying values of Read Threshold (Core, Fringe) 24

3.1 Distribution of dimers and pentamers across 50,000 reads sampled from the genome of Haemophilus Influenzae(Only a few distributions are shown). Distribution of dimers tends to Gaussian, two groups can be observed 30
3.2 Distribution of dimers and pentamers across 50,000 reads sampled from the genome of Haemophilus Influenzae(Only a few distributions are shown). Distribution of pentamers tends to Poisson, three groups are seen. 31
3.3 Illustration of a Two-way Poisson Mixture Model for Metagenomic Data. Each cluster represents a species and is modeled as a distribution of words comprising it. Each word follows a different distribution. However, not all words in a class have significantly different parameters. Therefore, the words can be divided into groups and words within the same group can be constrained to have identical parameters. 46
3.4 Our method converges for all cases tested and is robust to the choice of initial conditions. 47
3.5 Performance of Poisson Mixture Model vs. **Left.** Dinucleotide divergence. **Right.** Word length over 450 datasets with δ^* values ranging from 34 to 340 47

3.6 GMM stands for Gaussian mixture model (without word grouping). The top figure compares the performance of the three methods on 8 datasets (X-axis shows the abundance ratio and the species contained in the dataset.). The bottom figure plots the δ^* values for the corresponding datasets. The X-axis shows the corresponding read lengths. Here, the δ^* (measured on 50 kb contigs) ranges from 34 to 340. 48
3.7 Performance of Poisson Mixture Model at different reads lengths (50-1000 bp). Datasets with low δ^* values (100-150) were chosen. 49
3.8 Performance of Poisson Mixture Model at different coverage ratios in a 2-species datasets. Datasets with low δ^* values (100-150) were chosen. 49
3.9 Results on the Acid Mine Drainage Dataset 50

4.1 This figure illustrates the distribution of dimers and pentamers across 50,000 reads sampled from the genome of Haemophilus Influenzae. a) Distribution of dimers tends to Gaussian and is approximated by a Poisson, two distinct groups can be observed. b) Distribution of pentamers tends to Poisson, three groups can be observed. 52
4.2 Comparison of performance of Bayesian mixture of Poissons and Multinomials with their Naive Bayes counterpart for 2-species dataset with δ^* (measured on 50-kb contigs) values ranging from 34 to 340, with 34 corresponding to "closely similar species" and 340 to "very distant species". We used a word length of 4. 62
4.3 Performance of Bayesian mixture of Poissons a) Accuracy Vs. length of word b) Accuracy Vs. read length c) Accuracy Vs. number of clusters d) Accuracy Vs. coverage 63
4.4 Performance of Two-way Bayesian Poisson mixture model for values of word groups, L, varying from 10 to 1024. A word length of 5 is used. Each dataset contained 50,000 reads of 500 bps each . . 65
4.5 Comparison of performance of Bayesian mixture of Poissons with Scimm. We varied the δ^* values from 60 to 300. We used a read length of 200 bps and word length of 4. 66

5.1 Virus Haplotype model in 2-D sequence space. Variation within cluster is due to errors in sequencing. Variation between clusters is due to haplotype differences. 69

5.2 Virus population with two haplotypes in frequencies, x_A and x_B. Frequency spectrum of l-tuple counts in the population. The figure depicts two genomes of length G and a shaded region that corresponds to the base positions on which the two genomes differ (i.e., the suspect region). 73
5.3 Diagramatic illustration of Mutant-Bin 84
5.4 **Top**. Snapshot of 600 bp length of three genomes in a sample (with the crosses representing the derived nucleotides). The lowest genome is the designated reference. **Bottom**. Frequency spectrum of l-tuples count along the length of the genome for a viral population containing three haplotypes in the mixing ratio of 1:3:5. 85
5.5 Comparison of effectives of two different error correction techniques 1. Thresholding of l-tuples 2. Spectral Alignment 85
5.6 Estimated haplotype frequencies in sampling from mixing ratios (solid black lines) indicated beneath the panel for diversities 0.1-10%. 86
5.7 Estimated haplotype frequencies by ShoRAH in sampling from mixing ratios (solid black lines) indicated beneath the panel for diversities 0.1-10%. All parameters of ShoRAH were set to default values and the algorithm was run for 5000 iterations. 87
5.8 Precision**(Left)** and recall**(Right)** for datasets containing 2 haplotypes in the ratio shown to the left of the figures. X-axes shows the diversities varying from 0.1-10%. 87
5.9 F-measure with varying coverage. Parameters:Simulated datasets with two haplotypes in mixing proportion of 1:3, at diversities of 4-6% 88
5.10 F-measure with the number of haplotypes in the population. Number of reads = 20000. Haplotypes were considered with frequencies corresponding to 1:3:5:10:12. 89

6.1 Tree Construction: Using the frequency of differences of the viral genomes from the root of the tree, we can work backwards to reconstruct the evolutionary tree 96

List of Tables

2.1 Summary of the results of experiments for oligomer length = 6 and varying Read Thresholds. 25
2.2 Summary of results on Acid Mine Drainage and simulated dataset 27
2.3 Functions of symbionts in Termite Metagenome 27

3.1 Performance of Gaussian Mixture Model (without word grouping) on datasets containing more than 2 species, at various abundances on reads of length 500 bp. AR stands for Abundance Ratio . . . 41
3.2 Performance of Poisson mixture model on datasets for different values of L and word length of 5. Here, N.W.G stands for no word grouping. The maximum accuracy achieved is in bold. Each dataset contains 50,000 reads of length 500 bp. 42
3.3 Comparison of performance of Gaussian mixture model (GMM) with 2-way Poisson mixture model (PMM) for datasets with low δ^* values. Each dataset contains 50,000 reads of length 500 bp. . 42
3.4 Performance of Poisson Mixture Model (without word grouping) on datasets across various taxonomic ranks. Each dataset contains 50,000 reads of length 500 bp. AR stands for Abundance Ratio . 43

5.1 Effect of error correction using spectral alignment with thresholding on reduction in the number of erroneous l-tuples. Parameters: Simulated datasets with two haplotypes in mixing proportion of 1:3, at diversities of 4-6%, containing 10,000 reads. If we consider an error rate of 0.5% per bp, then approximately 4.9% of the l-tuples and 71% of the reads are expected to be contaminated with at least one sequencing error. 82

Acknowledgments

I would like to thank my mentor and adviser Dr. Raj Acharya, whose guidance and encouragement has been of immense value to me. He has been a great source of inspiration. I want to thank him for giving me the freedom to pursue my interests. I am grateful for his contribution of ideas and funding to make my Ph.D. experience productive and stimulating.

I would like to thank my co-advisor, Dr. Mary Poss, who has been a constant source of support throughout my thesis work. I am grateful to her for many insightful discussions and suggestions, which really helped me in understanding the biological aspects of my research.

I would like to thank Dr. Webb Miller for always being ever so approachable and available to my naive questions. His passion for research and work-ethic is both admirable and infectious. I would like to thank Dr. Padma Raghavan for her feedback on my research and thesis and for being a supportive committee member despite her busy schedule. I would like to thank Dr. Natalie Fedorova for introducing me to the field of metagenomics and contributing a great number of ideas to my research.

This thesis would have been impossible without the motivation of my parents who have instilled in me the desire to learn. I would like thank my brother Sandeep Prabhakara and his wife Tisha Agarwal for being there and providing support when I needed it the most. I am grateful for my family who raised me with a love for science and supported me in all my pursuits.

My time at Penn State was made enjoyable in large part due to the many friends and groups that became a part of my life. Thanks to Mohammed Habibulla and Hrishikesh Amur for their time, interest and help in reviewing my papers. Special mention to Pooja Nadkarni for being such a wondering roommate and a great cook. I would like to thank my labmates Aravindhan Venkateswaran, Ranjit Ganta and Raunaq Malhotra for numerous intellectual conversations and assisting in my research. Thanks also go to my fellow graduate students for making this department a great place to be and for having stuck it out in grad school

with me.

Dedication

To my parents,
Geeta and Prabhakara

Chapter 1

Introduction

1.1 Next Generation Sequencing

The recent introduction of next generation Sequencing (NGS) technologies capable of producing millions of DNA (Deoxyribonucleic acid) sequence reads in a single run is rapidly revolutionizing the landscape of genetics. NGS technologies generate data more efficiently, economically and with a greater depth than ever before [40, 48]. NGS has opened up an array of possibilities for many applications of which, the characterization of diversity of heterogeneous microbial environments has recently gained significant interest. It has empowered scientists with a tool to compare the Genome of many organisms at a time and look into minute differences that provide a new understanding of biology. Although a host of methods for whole genome Assembly have been developed, reconstruction of individual clones from heterogeneous populations such as metagenomes and viral quasispecies still remains a challenge. Compared to existing technologies, reads produced by NGS are typically shorter and more error-prone. The volume of sequence data acquired by environmental sequencing is several orders of magnitude larger than that acquired in single organism genomics. Thus, several computational challenges arise while analyzing deep sequence data from heterogeneous populations. The computational methods we present in this proposal aim at quantifying the microbial diversity within metagenomes and viral populations based on a set of deep sequencing reads.

1.1.1 Genomics

Genomics is a field of genetics concerning the study of genomes of a single organism. Each cell of a living organism contains a copy of the genome, which is a complete set of instructions for creating the organism. The genome is the entirety of organisms hereditary information. It is the blueprint for all cellular structures and functions of life. The biological information contained in a genome is encoded in its deoxyribonucleic acid (DNA). The DNA chain is mainly made of four repeating units called nucleotides or bases: adenine(A), thymine(T), guanine(G), cytosine(C). And it is the sequence of bases in the genome that specifies the exact genetic instructions that create the organism. Therefore, determining the entire sequence of bases in an organisms genome is instrumental in understanding the genome.

DNA sequencing is the process of determination of precise sequence of nucleotides in a DNA sample. The advent of DNA sequencing has significantly accelerated the pace of biological research and has been instrumental in sequencing the human genome, in the Human Genome Project[43]. Even after sequencing the genome, much work remains to be done. Scientists still need to decipher the string of letters to understand the functions of the various Gene that make the genome, how the genes are related, the regulatory regions that turn the gene on and off, as well as the stretches of nonsense DNA - after all, the genes account for less than 25% of the DNA in a genome.

How is the genome sequenced? The genome is sequenced in pieces. The whole genome cannot be sequenced at once because the chemical reactions used to decode the DNA are accurate for only up to 600-700 bases at a time. So instead, the process of sequencing begins by breaking several copies of the whole genome into millions of relatively short fragments. Then, the DNA sequencing machine determines the order of bases in each of these fragments. Next, a set of program called assemblers determine the overlaps between the fragments and reconstruct the original genome sequence. Much of the work involved in sequencing is aking to putting together a giant biological jigsaw puzzle, mostly without a reference. This general technique is known as shotgun sequencing and was pioneered by Frederick Sanger in 1982[59]. The technique took a major step forward in 1995, when a team led by Craig Venter, Robert Fleischmann and Hamilton Smith used

it for large scale sequencing of the 1.83 million base pair(Mbp) genome of the bacterium *Haemophilus influenzae*[22].

This seemingly simple process is beset with many technical challenges. Most notable of these is the occurrence of repetitive sections of DNA called repeats. The human genome, for example, includes some repeats that occur in more than 100,000 copies each. Similar to pieces of sky in jigsaw puzzles, reads belonging to repeats are difficult to position correctly. Further complicating assembly, is the presence of errors in data, from limitations in sequencing technology as well as from human mistakes during laboratory work. Moreover, some DNA fragments from each genome are impossible to sequence, resulting in gaps in Coverage.

1.1.2 Metagenomics

Metagenomics is defined as the study of genomic content of microbial communities in their natural environments, bypassing the need for isolation and laboratory cultivation of individual species[12]. Its importance arises from the fact that over 99% of the species yet to be discovered are resistant to cultivation[51]. This limitation imposed by cultivation of isolated clones has severely skewed our view of microbial diversity. Metagenomics promises to enable scientists to study the full diversity of the microbial world, their functions and evolution, in their natural environments.

Metagenomic projects collect DNA from environments that are characterized by large disparity in sequence coverage and species distribution. Sequencing technologies are then used to survey the metagenomic content. The recent ultra-high throughput sequencing technologies, such as Roche 454, Illumina/Solexa and ABI SOLiD produce short reads, 25-400 base pairs (bp), at a much higher coverage and considerably lowered sequence costs. The growth in the size of the datasets is fast outpacing the computational power needed to analyze it.

What is so challenging about Metagenomic Assembly? In single genome sequencing, we can be certain that all extracted DNA fragments belong to the same genome. This makes sequence assembly and annotation in a single genome project tractable, however, this is not the case for a metagenome. In majority of metagenomic samples, it is not possible to isolate and culture individual

clones from the metagenome. It is further complicated by the fact that the data comes from heterogeneous microbial communities, where the number of species as well as the relative abundance of each of these species is also unknown. In a metagenome study conducted by Venter *et al*[22], the DNA sequences sampled from Sargossa Sea indicated the existence of far more diverse microbial communities than previously thought. The length of each fragment can be anywhere between 20 base pairs(bp) and 700 bp, depending on the sequencing method used. Usually environmental sequence sampling produces very few species with a high abundance. Species with low abundance account for majority of the sample.

Many of these species do not have a fully sequenced genome available. Moreover, the sequence data is incomplete and fragmentary. Each fragment was obviously sequenced from a specific species, but in many cases it is impossible to determine the true species of origin. From clustering, to assembly, to annotation and function prediction, bioinformatics is posed with new challenges in handling noisy, huge and often partial sequence data. When a closely related fully sequenced genome is available, the standard approach is to use an alignment program to perform comparative assembly. But, most metagenomic sequences do not provide a significant match to the sequences in existing databases. In such cases, traditional overlap-layout-consensus methods cannot be used. More often than not, there is the danger of assembling sequences from different species, thereby creating inter-species chimeras. Phrap, Euler, AMOS, Forge, Arachne, Velvet and the Celera assembler are all assemblers that were developed for single genome assembly. A fundamental shortcoming of metagenomics is that only genomes of high abundance species can be assembled near completely. Lack of coverage makes it almost infeasible to obtain complete genomes of organisms from complex microbial communities. Short sequences that have been fragmented from their original genomes can be assembled to lengths usually not exceeding 5,000 bp, thereby making the reconstruction of a whole genome very difficult. Not only is the data fragmented and incomplete, but the sequence data acquired from environmental projects is several orders of magnitude larger than those from single species genomics[71]. Furthermore, metagenome datasets are beset with increased amounts of polymorphism and horizontal gene transfer. Sequences from closely related species will most likely have homologous sequences shared

between them, hindering their separation[10]. Moreover, the abundances of different species can be potentially skewed such that the within-species variance is overwhelming compared to the between-species variance[11]. We thus need to adapt traditional approaches to analyze metagenomic sequences.

For these reasons, computational biologists have been developing new algorithms to analyze metagenomic sequences. The computational challenges put forth by metagenomics are new and very exciting. We are entering an era similar to the first genomic revolution. An additional analytics step in metagenomics that is not required in single genome assembly, is binning the reads belonging to a species i.e. the need to associate the reads with its source organism. Clustering methods aim to identify the species present in the sample, classify the sequences by their species of origin and quantify the abundance of each of these species. Until now, methods to classify short sequences generated by the ultra high throughput sequencing technologies have been relatively inaccurate. It is critical that we develop fast and accurate tools to address the challenges posed by the nature of metagenomic data. Clustering of metagenome reads is one such tool that provides deeper insight into the structure of the community. It can lead to faster and more robust assembly by reducing the search space[44]. Clustering and assembly statistics can be used to model the ecological and population parameters.

1.1.3 Heterogeneous Viral Populations

High genetic variability in viral populations plays an important role in disease progression, pathogenesis and drug resistance. The last few years has seen significant progress in the development of methods for reconstruction of viral populations using data from NGS technologies. These methods identify the differences between individual Haplotype by mapping the short reads to a reference genome. Much less has been published about resolving the population structure when a reference genome is lacking or is not well-defined, which severely limits the application of these new technologies to resolve virus population structure.

At any given time, within-host virus populations consist of a collection of distinct, albeit closely related genetic variants, known as quasispecies. Each indi-

vidual variant, a haplotype, occurs with a different relative frequency. The high genetic diversity of a pathogen population has important consequences in disease progression as it allows the virus to respond to changes in the host environment such as evading host defenses and therapeutic interventions. The high coverage and enormous sequence data output by NGS technologies have the potential to resolve the genetic variation within a virus sample and thereby infer the population structure and composition, which will directly benefit research on disease progression, drug resistance, vaccine design and viral evolution[25].

1.2 Research Contributions

Clustering methods aim to identify the species present in the sample, classify the reads by their species of origin and quantify the abundance of each of these species. Clustering provides deeper insight into the structure of the community. The efficacy of clustering methods depends on the number of reads in the dataset, the read length and relative abundances of source genomes in the microbial community. In this context, in my dissertation research, I will focus on development of methods to characterize and analyze the taxonomic content of vast amounts of short metagenome reads and heterogeneous viral populations. The main contributions of the dissertation are: (i) A two-pass soft clustering method that is a hybrid of comparative and composition based methods. (ii) An unsupervised naive Bayes mixture model based on Poisson or Gaussian distributions to model reads within each bin. (iii) An unsupervised multivariate Bayesian mixture model based on Poisson and Multinomial distributions for clustering discrete sequence data. (iv) A computational framework, called Mutant-Bin, for clustering individual haplotypes in a heterogeneous population and determining their prevalence, based on a set of deep sequencing reads. A brief explanation of each of the methods is provided below.

1.2.1 Clustering Algorithms for Metagenomics

SIMCOMP-A Hybrid Soft Clustering Algorithm: Metagenome reads are characterized by increased amounts of polymorphism and horizontal gene trans-

fer. Reads from closely related species will most likely have homologous sequences shared between clusters that occlude the cluster boundaries[10]. Moreover, the incomplete and fragmentary nature of the metagenome reads reduces the quality of annotation.

In light of the new data, we propose a two-pass semi-supervised algorithm for fuzzy clustering of metagenome reads that is a hybrid of comparative and composition based approaches. This method significantly reduces the size of the metagenome dataset while maintaining an accurate representation of its functional and taxonomic content. Overlapping clusters generated by a fuzzy clustering algorithm elegantly handle the problems associated with the nature of metagenomic data while providing tolerance for the noise in the data due to errors in sequencing and fragmentation. Our primary goal is to enrich the dataset into a small number of clusters such that reads within a cluster are phylogenetically closer than reads from different clusters. In our method, the comparative analysis of reads avails a priori biological knowledge in the existing database to form an initial set of seeded clusters. In the following pass, the composition based characterization of the remaining fraction of reads into existing clusters, facilitates a means of exploring novel species. The secondary goal is to identify polymorphic and conserved regions and capture them within the soft boundaries of the clusters. Due to evolution, the nucleotide composition of genomes belonging to same lower taxonomic levels can be very similar. Regions with overlapping clusters capture reads that are phylogenetically closer. Our two-step algorithm is as follows:

- In the first pass, a comparative analysis of the metagenome reads against an existing database, using BLAST (Basic Local Alignment Search Tool)[2], extracts reference sequences from within the dataset to form an initial set of seeded clusters. Reads that have a significant match to the database are clustered by their phylogenetic provenance.

- In the second pass, the global clade-specific characteristics (e.g. oligomer frequency) are used to cluster the remaining reads by a fuzzy possibilistic leader clustering algorithm described in[58]. Our algorithm groups the reads into overlapping clusters. Each cluster is defined by a core consisting of

reads that definitely belong to the cluster and a fringe that has reads which may overlap with other clusters (representing homologous sequences). The fringes of the clusters accommodate the ambiguity associated with reads in the dataset. The resulting cluster leaders can be used as an accurate estimate of the phylogenetic composition of the metagenomic dataset.

The significantly reduced size allows a compact yet comprehensive overview of the dataset. The proposed method does not require assembled contigs or training on a reference set, nor does it make any assumptions on the number of species or the nature of the dataset. It makes use of a reference database, however is not dependent on it. Our method enriches the dataset into a small number of clusters, while accurately assigning fragments as small as ~ 100 base pairs. An important consequence of our method is that the fuzzy boundaries between clusters capture the misplacements of reads due to over representation of conserved regions, without clipping potentially useful sequences. In Chapter 2, we present the algorithm in detail and discuss the experimental results on two datasets: a simulated dataset of 454 reads that are 100 bps at varying coverage, and acid mine drainage metagenome dataset.

Naive Bayes Mixture Model: In Chapter 3, we formulate an unsupervised naive Bayes multi-species, multi-dimensional mixture model for reads from a metagenome. We use the proposed model to cluster metagenomic reads by their species of origin and to characterize the abundance of each species. Recent studies in metagenomics indicate the presence of a "genome signature", a compositional parameter which reflects the relative abundance of different words along a genome that can be used to distinguish between reads from different species. We model the distribution of word counts along a genome as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare. We employ either a mixture of Gaussians or mixture of Poissons to model reads within each bin. An additional reason to use these distributions is their flexibility and ease of parameter estimation. Such a paradigm characterizes the compositional heterogeneity of the words along a genome, signifying its genome signature. Further, we handle the high-dimensionality and sparsity associated with the data, by grouping the set of words comprising the reads, resulting in a two-way mixture model. Finally, we derive an unsupervised Expectation Maximization algorithm for the models. Our

method provides a general statistical framework for modeling metagenome reads. We demonstrate the accuracy and applicability of this method on simulated and real metagenomes. Our method can accurately cluster reads as short as 100 bps and estimate the species abundance as well. Our method outperforms LikelyBin, another unsupervised composition-based binning method for metagenomes, on datasets of varying abundances, divergences and read lengths.

Naive Bayes is the simplest Bayesian network that does not represent any variable dependencies. Typically, even if the sequences of based in a DNA are independently and identically distributed, distribution of word counts is not independent due to overlaps. Though, in practice, methods for exact inference are often computationally expensive. The assumption of independence between the words in a read, makes the otherwise complicated problem tractable. Naive Bayes takes time linear in the number of components.

Bayesian Mixture Model: We present an efficient multivariate Bayesian mixture model based on Poisson and Multinomial distributions for clustering discrete sequence data. The structure of Bayesian networks efficiently encodes the conditional dependencies between the words due to overlaps. The Poisson mixture model is derived from the assumption that the distribution of word counts along a genome follows a Poisson distribution. The Multinomial mixture model is derived as a standardized Poisson mixture model. We present a two-way clustering approach to handle the high-dimensionality and sparsity associated with the data. It combines the words along the reads into word groups and then constrains the parameters for words within the same group to be identical. The motivation of this method is to overcome the bottleneck of Naive Bayes by taking into account the conditional dependencies between the word counts within the reads. We will use the Bayesian networks to specify the structure of the network and learn the parameters as well.

Our method can cluster reads as short as 50 bps with accuracy over 80% and estimate species abundance as well. The Bayesian mixture of Poissons and Multinomials outperform their Naive Bayes counterparts on datasets of varying abundances, divergences and read lengths. Our method is robust to the number of species in the dataset, read lengths and relative abundances of source genomes in the metagenome. Despite our specific application to metagenomics, the Bayesian

mixture models are useful for classifying any high-dimensional discrete sequence data.

1.2.2 Haplotype Estimation of Viral Populations

We describe a computational framework, called Mutant-Bin, for clustering individual haplotypes in a viral population and determining their prevalence, based on a set of deep sequencing reads. The main advantages of our method are that: (i) it enables determination of the population structure and haplotype frequencies when a reference genome is lacking; (ii) the method is unsupervised; the number of haplotypes does not have to be specified in advance; (iii) it identifies the polymorphic sites that co-occur in a subset of haplotypes and the frequency with which they appear in the viral population. The method was evaluated on simulated reads with sequencing errors and 454 pyrosequencing reads from HIV (Human Immunodeficiency Virus) samples. Our method clustered a high percentage of haplotypes with low false positive rates, even at low genetic diversity.

1.3 Organization

The remainder of the dissertation is organized as follows. Chapter 2 discusses a hybrid clustering method SIMCOMP for metagenomes in detail. The Naive Bayes approach to clustering metagenome reads based on Gaussian and Poisson distribution of word counts is discussed in Chapter 3. Chapter 4 presents an alternative approach, the two-way Bayesian mixture model that overcomes the deficiencies of the Naive Bayes methodology. In Chapter 5 we present a computational framework called Mutant-Bin for clustering individual haplotypes in a viral population and determining their prevalence, based on a set of deep sequencing reads. Chapter 6 concludes this dissertation with a summary of contributions and future directions for research. The software programs are publicly available and may be accessed at http://www.cse.psu.edu/ sap263/software.html.

Chapter 2

SIMCOMP: A Hybrid Clustering Algorithm

A major challenge facing metagenomics is the development of tools for the characterization of functional and taxonomic content of large datasets containing short metagenome reads. In this chapter, we present a two pass semi-supervised algorithm, SimComp, for soft clustering of short metagenome reads. SimComp is a hybrid of comparative and composition based methods. The proposed method significantly reduces the size of the metagenome dataset while maintaining an accurate representation of its taxonomic and functional content. In the first pass, a comparative analysis of the metagenome reads against a database using BLASTx extracts the reference sequences from within the metagenome to form an initial set of seeded clusters. Those reads that have a significant match to the database are clustered by their phylogenetic provenance. In the second pass, the remaining fraction of reads are characterized by their species-specific composition based characteristics. SimComp groups the reads into overlapping clusters, each with its read leader. We make no assumptions about the taxonomic distribution of the dataset. The overlap between the clusters elegantly handles the challenges posed by the nature of the metagenomic data. The resulting cluster leaders can be used as an accurate estimate of the phylogenetic composition of the metagenomic dataset. Our method enriches the dataset into a small number of clusters, while accurately assigning fragments as small as 100 bps.

2.1 Related Work

The last decade has seen an explosion in the number of computational methods developed to analyze the metagenomic data. Literature abounds in methods for classifying (as opposed to clustering) metagenome reads into taxon-specific bins [42, 7, 44]. Current approaches to metagenomics binning can be classified into two main categories: similarity based and composition based.

The similarity-based approaches align the reads to close phylogenetic neighbors and hence depend on the availability of closely related genomes in existing databases[23, 17, 27]. MEGAN, a metagenome analysis software system [27] is a representative example of this kind. It uses sequence homology to assign reads to common ancestors based on best match as given by BLAST (Basic Local Alignment Search Tool)[2]. As most of the extant databases are highly biased in their representation of true diversity, such methods fail to find homologs for reads derived from novel species.

A second class of computational methods bin the reads based on DNA composition. It is this class of methods that is of interest to us. These methods rely on the intrinsic features of the reads such as oligonucleotide distributions[42, 7, 11, 10, 33, 32], codon usage preference[4] and GC composition[6] to differentiate between reads belonging to different species. These "genome signatures" are known to be fairly constant throughout the genome. The underlying basis is that the distribution of words in a DNA is specific to each species and undergoes only slight variations along the genome. By establishing the dictionary of words used by a species and their frequency of occurrence, one can point out the basic words of the genome[19]. A significant limitation of most composition-based methods developed so far is that they do not perform well on reads shorter than 500 bps. Methods for composition-based clustering of metagenome reads complements those based on similarity.

Phylopythia [42] is a supervised composition-based classification method that trains a support vector machine to classify sequences of length greater than 1 kbps. Phymm uses interpolated Markov models to characterize variable length DNA sequences by their phylogenetic group [7]. Its accuracy of assignment drops drastically (to just 7.1% at Genus level) for short reads and reads from unknown

species. Nasser *et al.* [44] demonstrated that a k-means based fuzzy classifier, trained using a maximal order Markov chain, can separate fragments that are about 1 kbps long at the Phylum level with a high accuracy. Rosen *et al.* trained a Naive Bayes classifier using publicly available microbial genomes[56]. CompostBin is a semi-supervised algorithm for grouping fragments that uses a novel weighted PCA (Principal Component Analysis) and a normalized cut clustering algorithm to classify the sequences[11]. They have demonstrated an error rate bounded by 10%, when guided by information from phylogenetic markers, on datasets of low complexity. However, the accuracy of this method on reads less than 1 kbps has not been shown. Li *et al.* proposed a composition based leader clustering algorithm that clusters highly homologous sequences in order to condense a large database using word-filtering[37]. Recently, Chan *et al.* developed a semi-supervised seeded growing self-organizing map (S-GSOM) [10] to cluster metagenomic sequences. It extracts 8-13 kbps of flanking sequences of highly conserved 16S rRNA from the metagenome and uses them as seeds to assign the remaining reads using composition-based clustering. The caveat with SOMs is that it was shown to work well only on DNA fragments that are longer than 8kbps and lose much accuracy for reads with length below 1 kbps. All the above supervised methods depend on the availability of reference data for training. A metagenomic dataset, may however, contain reads from unexplored phyla which cannot be labeled into one of the existing classes. The accuracy of these methods on dataset containing reads from unknown species is yet to be demonstrated.

LikelyBin is an unsupervised method that clusters metagenomic sequences via a Monte Carlo Markov Chain approach[33]. The method was tested on samples that were sufficiently divergent according to derived criteria. Scimm is a recently developed state-of-art model-based approach to sequence clustering where interpolated Markov models represent clusters and optimization is performed using a variant of the k-means algorithm[32]. In the next few chapters, we compare the accuracy of our proposed methods with LikelyBin and Scimm on datasets of different divergences. Abundance Bin can be used to classify reads from species with different abundance levels[72]. Recall that coverage is usually incomplete in a metagenome. Abundance Bin does not perform well on reads from species that do not differ much in their abundance levels. However, if it is known a priori that

the reads differ widely in their abundances, then we recommend using Abundance Bin to separate organisms by their varied coverages. Most metagenomic analysis methods until now have been relatively inaccurate in classifying short reads. Poor performance on the short fragments is mostly due to the high dimensionality and sparsity associated with the data. Moreover, the abundances of different species can be potentially skewed such that the within-species variance overwhelms the between-species variance[11].

2.2 Background and Motivation

As we have seen before that methods for clustering reads proposed so far in literature can be categorized into two main approaches; comparative(or similarity) and composition based. Comparative based methods align metagenomic sequences to close phylogenetic neighbors in existing databases and hence depend on the availability of closely related genomes in the database[23, 17, 27]. Such methods fail to find any homologs for new families. Composition based methods, on the other hand, distinguish between clades by using intrinsic features of reads such as Oligomers frequencies[42, 7, 64], codon usage preferences[4] or GC content[6]. The strength of this approach is that no reference database is required. However, oligomer composition of reads shorter than 1 kbps carry insufficient signal to be able to differentiate between species. Composition based clustering of metagenome reads complements the comparative analysis[7].

Supervised methods depend on the availability of reference data for training[7, 42, 44]. These methods assume the prior knowledge of the number of classes. A metagenomic dataset may contain reads from unexplored phyla which cannot be labeled into one of the existing classes. As most of the extant databases are highly biased in their representation of true diversity, comparative methods such as MEGAN fail to find any homologs for new families. Most metagenomic analysis methods until now have been relatively inaccurate in classifying reads as short as 100 bps.

Increased amounts of polymorphism and horizontal gene transfer in reads of a metagenome, lead to conflicts in assembly and taxonomic analysis. Reads from closely related species will most likely have homologous sequences shared

between clusters that fuzzify the cluster boundaries[10]. Another characteristic of these datasets is the incomplete and fragmentary nature of the metagenome reads that reduces the quality of annotation. However, clipping low quality reads such as chimeras can exclude potentially useful sequences. Hence, in light of the new data, we need to adapt the traditional approaches to metagenome analysis. Overlapping clusters generated by a soft clustering algorithm such as the one proposed in this chapter elegantly handle the problems associated with the nature of metagenomic data while providing tolerance for the noise due to errors in sequencing and fragmentation. The soft boundaries between clusters provide the flexibility to capture the misplacements of reads due to polymorphism or over representation of conserved regions, thereby providing interesting insights into the data.

Our work is inspired by the works of Dalevi *et al.*[17] and Folino *et al.*[23]. In [17], the authors propose a method for clustering reads based on a set of Protein, called Proxygene. The protein hits are obtained by BLASTx (specialized nucleotide-protein BLAST) of the reads against a reference proteome database. Their work is extended in [23], where a method based on weighted proteins is used to cluster the reads, resulting in overlapping clusters, each represented by a proxygene. The underlying basis of the above methods is that a high sequence similarity between the read and the proxygene implies phylogenetic proximity of the organisms from which they originated [17]. Consequently, the taxonomic annotation of the proxygene can be used in assessing that of the reads in the cluster. Both the methods use the comparative approach and hence rely on the use of a reference database that contains closely related genomes. However, in a typical metagenome dataset, majority of the reads may exhibit no similarity to any known sequence in the database. In such a scenario, these methods will fail to assign these reads to any cluster.

In this chapter, we propose a two pass semi-supervised algorithm for soft clustering of short metagenome reads. We call our method SimComp; a hybrid of similarity and composition based methods. The objective of our method is to enrich the dataset into a small number of clusters such that reads within a cluster are phylogenetically closer than reads from different clusters. Each cluster is defined by a core consisting of reads that definitely belong to the cluster

and a fringe that has reads which may overlap with other clusters. We make no assumptions about the taxonomic distribution of the metagenome dataset. SimComp makes use of a reference database, however is not dependent on it.

In the first pass, a comparative analysis of the metagenome reads against an existing database, using BLASTx, extracts reference sequences from within the dataset to form an initial set of seeded clusters. Reads that have a significant match to the database are clustered by their phylogenetic provenance. In the second pass, the global clade-specific characteristics(e.g. oligomer frequency) are used to cluster the remaining reads by a soft leader clustering algorithm described in [58]. Our algorithm groups the reads into overlapping clusters, each with its read leader. The fringes of the clusters accommodate the ambiguity associated with reads in the dataset. It automatically performs the selection of the number of clusters. Essentially, the comparative analysis of reads avails a priori biological knowledge in existing protein database to form initial set of seeded clusters. Then, the composition based characterization of remaining fraction of reads, thereby facilitating a means of exploring novel species. Our method significantly reduces the size of the dataset, while maintaining an accurate representation of its functional and taxonomic content.

2.3 Methods

SimComp is based on the Adaptive Rough Fuzzy Leader Clustering presented by Asharaf *et al.*[58]. The authors use rough set theory to define the clusters. Each cluster has a core(lower bound) and a fringe(upper bound) and is represented by a read leader. The core contains all the reads that definitely belong to the cluster. Reads in the core are mutually exclusive between the clusters. There can be an overlap in the fringes of two or more clusters.

2.3.1 Comparative Clustering

In the comparative pass of the algorithm, as in [23, 17], we associate a list of protein hits with each read, identified by BLASTx. Each hit consists of one protein, two score values called bits and identities which describe the significance

of read-protein alignment, and a confidence value called E-value which describes the likelihood that the sequence will occur in the database by chance. We further use the measure defined in [23], explained below, for assigning weights to the each of the proteins, such that proteins that cover more reads are assigned smaller weights. Proteins that are below a predefined protein threshold form the proxygenes, the rest are discarded. The proxygenes are clustered with the corresponding best hit reads(as identified by BLASTx). For each cluster thus formed, the most representative read is chosen as the leader(seed of a cluster).

As in [23], from each hit that BLASTx outputs for a given read r, we extract a 4-dimensional vector $h = (p; S_B; Id; E)$ where p is the matched protein, S_B the bit score, Id the identities score, and E the E-value of that match. For a read r let Hit_r be the sequence, sorted in increasing order of E-values, of its hits. Denote by $r_1, ..., r_m$ the set of reads r with non-empty Hit_r. Let $P = \{p_1, ..., p_n\}$ be the set of proteins occurring in $\cup_{i=1}^{m} Hit_i$. For each protein $p \in P$, the set H_p is defined as:

$$H_p = \{h \in \cup_{i=1}^{m} Hit_i | h(1) = p\} \tag{2.1}$$

where $h(1)$ denotes the first component of the hit vector h. Thus H_p consists of the selected hits containing p. We use the equation described in [23] to assign weights to the each of the protein hits that BLASTx outputs. Weight of protein p is defined as:

$$w_p = 1 + \lceil \frac{1}{|H_p|} \sum_{h \epsilon H_p} (100 \frac{max\_score - S_B(h)}{max\_score - min\_score} + 100 - Id(h)) \rceil \tag{2.2}$$

where H_p consists of hits containing p, $S_B(h)$ and $Id(h)$, the bit and identity score of hit h respectively. For further details, we refer the reader to [23].

2.3.2 Composition Based Clustering

The reads remaining after the first pass are clustered using the soft leader clustering algorithm based on sequence composition. In this pass, each unclustered read is compared with the existing read leaders. The similarity between the read and the leaders along with the sequence thresholds determines whether the read gets added to the core of some cluster or fringes of one or more clusters, or the

read itself gets added as a leader. The steps in SimComp are outlined below.

2.3.3 Definitions

Cluster. Each cluster consists of a read leader, representative of the set of reads in the cluster. A cluster is defined by the following parameters:

- Protein threshold (PT): Proteins with weight below the threshold form proxygenes. Each proxygene is representative of a cluster with the corresponding reads(as identified by BLASTx). Rest of the proteins are discarded. The weight assigned to a protein is measured by two score values, i.e. bits and identities, and a confidence value called E-value[23].
- User defined core and fringe sequence similarity threshold for clusters (RT_C and RT_F): If the similarity between the read and its nearest leader is greater than RT_C, the read is added to the core of a cluster. Otherwise, if the similarity between the read and the corresponding cluster leaders is greater than RT_F, the read is added to the fringes of one or more clusters.

Sequence similarity. Each read is represented by a vector of oligomer frequencies, $v = (f_1, f_2...f_q)$; where for each oligomer of length n, $O = (o_1, o_2...o_q)$ is the set of all possible oligomers, f_i is the frequency of oligomer pattern o_i in the read, q is the number of oligomer patterns of length n possible, i.e. 4^n. Each vector is normalized relative to the length of the sequence. $S(x, y)$ gives the similarity between read x and leader y. We define sequence similarity as the number of fixed length oligomers shared between x and y.

Fuzzy membership. U_{ik} is the fuzzy membership of the read r_i in a cluster represented by Leader L_k.

$$U_{ik} = \sum_{j=1}^{N} \frac{S(r_i, L_k)}{S(r_i, L_j)} \tag{2.3}$$

2.3.4 SIMCOMP : Outline of the Algorithm

The algorithm proceeds in two passes. Let $R = (r_1, r_2, ...r_n)$, be the set of all reads and N be the number of clusters at any point in the algorithm.

I. Comparative Clustering: In the first pass, metagenome reads are grouped into clusters based on similarity of the reads to the proteins in the reference database.

1. Extract all proteins that R has hits to(by BLASTx).
2. Assign weights to all the proteins based on equation described below. Proteins with weight below PT form proxygenes. Weight of protein p is defined as:

$$w_p = 1 + \lceil \frac{1}{|H_p|} \sum_{h \epsilon H_p} (100 \frac{max\_score - S_B(h)}{max\_score - min\_score} + 100 - Id(h)) \rceil \quad (2.4)$$

   where H_p consists of hits containing p, $S_B(h)$ and $Id(h)$, the bit and identity score of hit h respectively (each hit is a 4-dimensional vector $h = \{p, S_B, Id, E\}$). For more details, we refer the reader to [23].
3. Each proxygene, along with the corresponding best hit reads (identified by BLASTx) form a cluster.
4. For each of the clusters, find a read leader that is most representative of the reads in the cluster, i.e. one whose sum of sequence similarity from all the other reads in the cluster is maximum.

II. Composition Based Clustering: In the second pass, we use the similarity measure based on oligomer frequency(defined above) to cluster the remaining reads.

1. All the reads from the original dataset that have not yet been clustered form the remaining read set. For each read in the remaining read set, compare the read with the existing read leaders. Depending on the value of RT_C, RT_F and sequence similarity between the read and the leaders, one of the three cases can arise for assignment of the current read:

(a) It gets added to the core of a cluster. The current read gets added to the core of a cluster represented by leader L_p, if :

$$\max(S(r_i, L_k)/k = 1...N) = D_{ip} \text{ and } D_{ip} > RT_C \tag{2.5}$$

(b) It gets added to the fringes of one or more clusters. r_i falls into the fringes of all the clusters L_p for which $S(r_i, L_p) > RT_F$ and $S(r_i, L_p) < RT_C$.

(c) Otherwise, r_i gets added as leader since it is outside the region defined by any of the existing clusters.

2.4 Results

We implemented our algorithm in Matlab. All experiments were run on an IBM X3550 server with 8GB memory. We tested our method on simulated metagenome datasets M1, M2 and M3, introduced in [17], each at different coverage levels (0.1X, 1X, 2X and 4X per genome). Reads from 22 genome projects were sequenced at Joint Genome Institute using the 454 pyrosequencing platform that produces ∼100 bps reads and split into three groups based on their phylogeny and number of reads to ensure similar sizes for the simulated datasets. We present results from experiments on M1 dataset here. The characterization of reads at the taxonomic level of an organism for M1 is as shown in Figure 2.1. We used the default parameters of BLASTx, and NR (Non-Redundant) protein sequence database as our reference. We have conducted experiments for varying values of user-defined thresholds(RT_C, RT_F) and lengths of oligomers. Based on the evaluation of our method on M2 and M3, we observed that proteins with weight below the 1^{st} percentile cover all the taxonomies that reads belong to. Therefore, we selected the 1^{st} percentile of weight as our protein threshold. The most time consuming component of SimComp is generating the BLASTx output. Once this output has been generated, the algorithm performs a single pass over the BLASTx output and the dataset to cluster the reads and hence is very efficient.

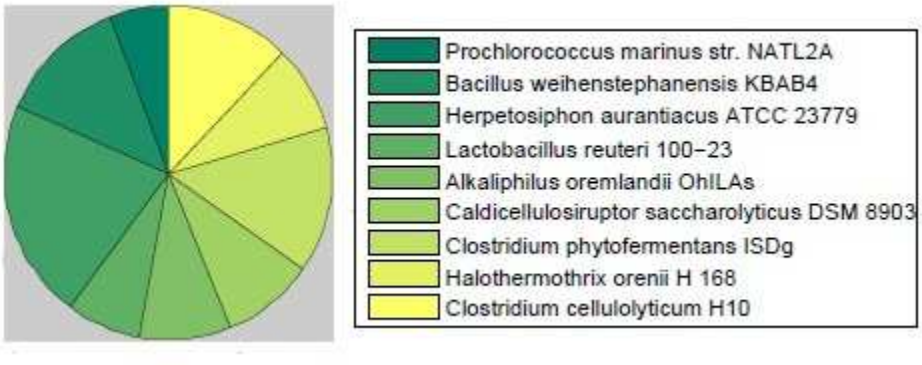


Figure 2.1. Organism level characterization of M1 dataset

2.4.1 Accuracy across Taxonomic Ranks

In this chapter, we use two measures to evaluate the effectiveness of our method: Mode Cluster Purity and Leader Cluster Purity. Mode Cluster Purity is defined as the maximum fraction of reads in a cluster belonging to the same taxon[23]. We define Leader Cluster Purity as the fraction of elements in the cluster belonging to the same taxon as the read leader. This measure determines how well our algorithm models the problem of classifying reads from species that have never been seen before. Depending on the elements of the cluster that we evaluate on, cluster purity can be further divided into core cluster purity(all the reads in the core of the cluster) and total cluster purity(all the reads in the cluster). In evaluating both the measures, we take into account only the non-singleton clusters, as a singleton cluster has a cluster purity of 1.

In Figure 2.2, we plot the taxonomic distribution of reads in M1 at phylum, class, order and family level($RT_C = 15$ and $RT_F = 12$ and length of oligomer = 6) as predicted by our algorithm. To measure the taxonomic distribution, all the reads in the cluster are assigned the same taxa as the read leader. Our method yields satisfactory results at all ranks. Hence, leaders of the clusters can be used as an accurate estimate of the phylogenetic composition of the metagenome. In [17, 23], only those reads that have significant hits in the BLASTx output are selected for further clustering, the remaining reads are discarded. As opposed to this, in our method, we cluster all the reads in the dataset, even if no significant hits to the reference database are obtained. In Figure 2.3, we have plotted three

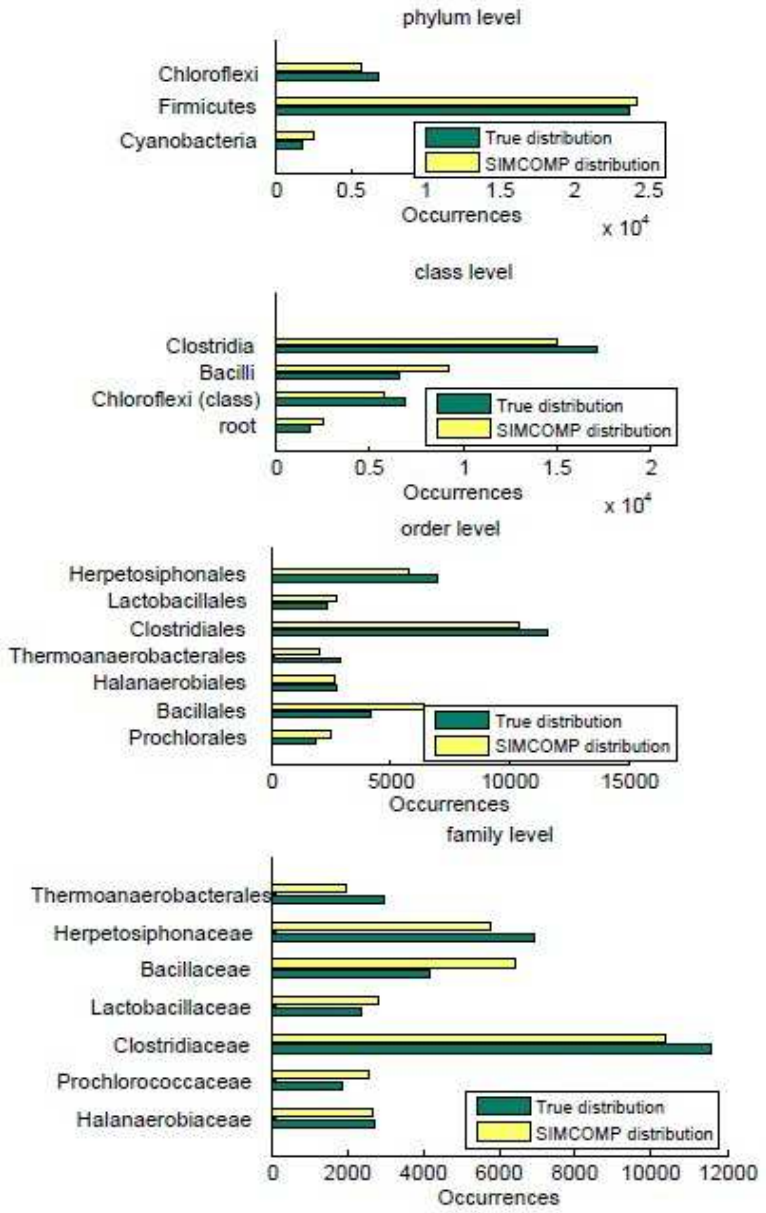


Figure 2.2. Taxonomic Distribution Across Ranks (Phylum, Class, Order, Family

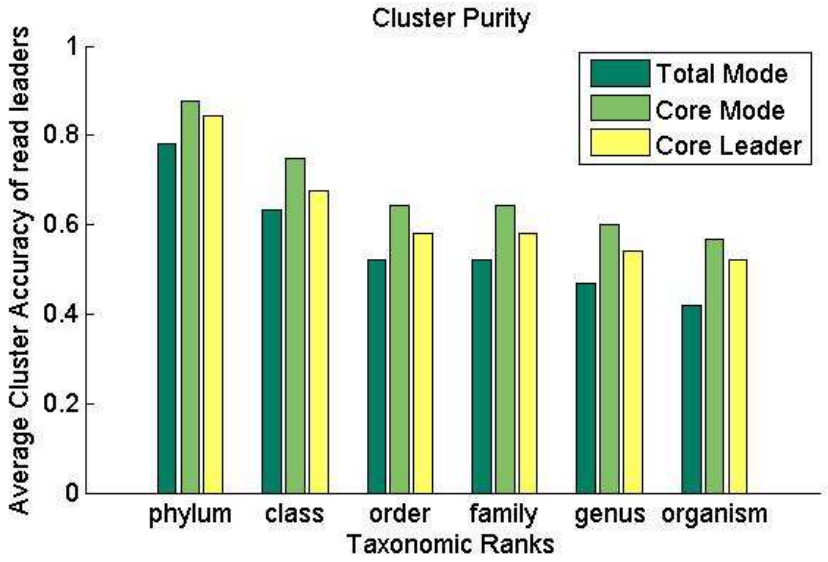


Figure 2.3. Average cluster purity across taxonomic ranks for ($RT_C = 15$ and $RT_F = 12$ and length of oligomer = 6, Number of Clusters = 2430)

measures for dataset M1 across all taxonomic ranks. By definition, mode cluster purity is greater than or equal to leader cluster purity. From the plot, we conclude that the cluster purity of the core is higher than that of the entire cluster at all ranks. This asserts our algorithms ability to filter out low quality reads into the fringe of a cluster.

2.4.2 Length of Oligomer

Oligomer frequency of genomes has been shown to reflect clade-specific characteristics and thus form a genome signature[31]. Teeling *et al.*[63] have shown that tetranucleotide frequency has a higher discriminatory power than GC content for phylogenetic grouping of reads. We have evaluated the accuracy of assignment of reads to clusters for a range of oligomers varying from trimers to hexamers. Figure 2.4 shows the plot of percentage of non-singleton clusters with purity values in the range [0.1,1] for varying lengths of oligomer. From our experiments, we conclude that hexamers have the best discriminatory power for clades at higher taxonomic ranks. With reads as small as 100 bps, not many reads cross that high a similarity threshold for hexamers. This explains the increase in number of singleton clusters with the increase in read threshold.

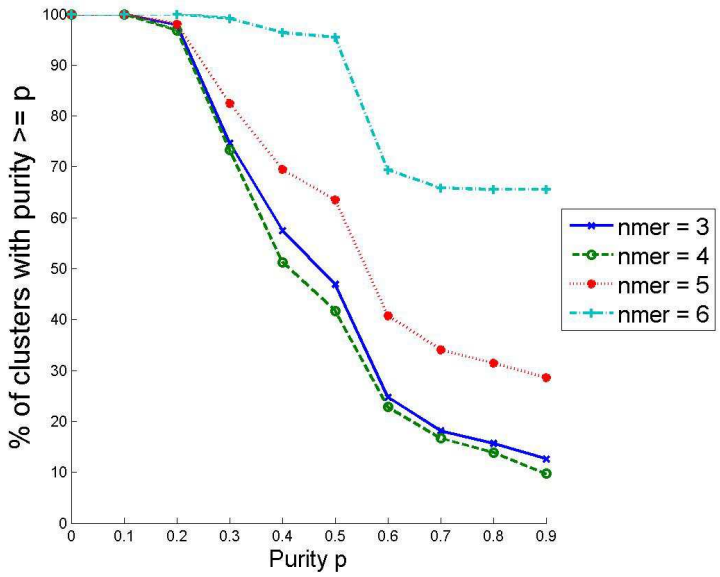


Figure 2.4. Plot of percentage of non-singleton clusters for different values of purity with $RT_C = 25$ and $RT_F = 22$ and varying values of oligomers

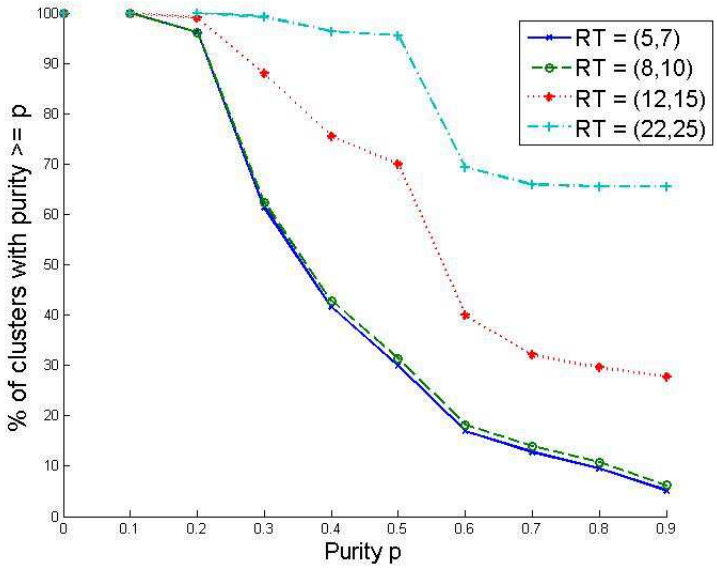


Figure 2.5. Plot of percentage of non-singleton clusters for different values of purity with oligomer lenght = 6 and varying values of Read Threshold (Core, Fringe)

Table 2.1. Summary of the results of experiments for oligomer length = 6 and varying Read Thresholds.

RT_C RT_F	10 8	15 12	20 17
Number of Clusters	1482	2430	14250
Maximum size of clusters	320	415	288
Number of singleton clusters	6	67	5865
Reduction factor	0.042	0.068	0.4
Mode Cluster Purity at Phylum level	79.93	88.14	96.95
Mode Cluster Purity at Organism level	40.88	61.75	88.41

2.4.3 Read Threshold

In our method, sequence similarity between two reads is measured as a function of number of fixed length oligomers shared between the two reads. A read is added to the core of an existing cluster only if the read similarity between the read and the cluster leader is above a certain threshold. Figure 2.5 plots the mode cluster purity for different values of read thresholds. The curve for $RT_C = 25$ clearly dominates the others. This is justified as clusters with large read thresholds are smaller in size and hence are likely to have a high purity. Table 2.1 summarizes the results for a fixed oligomer length of 6 and varying read thresholds. Cluster purity increases with the increase in read thresholds, for the reasons cited above.

2.5 Termite Metagenome

We analyzed the complementary DNA reads of Coptotermes Formosanus for the its phylogenetic characterization. We used SimComp to determine the taxonomic content of the meta-transcriptome sample and to determine the relative abundances of different species in the termite digestome. We set the identity threshold at 30% and E-value threshold at 1.0. The termite digestome can be defined as a pool of genes, from termite and its symbionts, that contribute to lignocellulose depolymerization and digestion.

Termites, formerly known as isoptera, are close relatives of cockroaches. Termites consist of seven families assigned across 2700 species, which are divided into

lower and higher termites. Six of the seven families are called lower termites, they possess flagellated protists and prokaryotes in their hindguts and are expected to provide useful genes for wood decomposition and fermentation. On the other hand, members of the remaining family Termitidae are called higher termites that contain approximately 85% of all known genera. Compared to lower termites that primarily feed on wood, higher termites show diverse feeding habits. Our reads come from the RNA (Ribonucleic acid) of Coptotermes Formosanus, which belongs to family Rhinotermitidae. To date, none of these species have been sequenced.

The termite gut exhibits one of the most complex microbial communities, consisting of diverse microorganisms from all three domains of life: Bacteria, Archaea and Eukarya. Majority of the reads were classified as Arthropoda suggesting maximal presence of host DNA. The most abundant host species was Drosophila. A plausible explanation is that, to date, no genome of termite or cockroach has been sequenced and the closest sequenced relative of termite is Drosophila. Species identification in termite gut turned out to be challenging as only a small fraction of symbionts diversity is obtained in the reads. Symbiont diversity comprised less than 15 % of the total reads. As the cDNA library was constructed from Poly(A)+ RNA, the bacterial population might be underrepresented in the dataset. Among the bacteria, the most frequently recovered phylotypes belonged to Bacteroidetes, proteobacteria and Firmicutes. The methanogenic archaea are permanent residents in the termite gut and a major source to natural methane on earth [66]. A variety of fungi were also identified in the termite gut. These fungi are known to play a significant role in the emergence of social homeostasis in the termite colonies. The fungi are part of an extracorporeal digestive system that converts undigested woody material in plants into higher quality oligosaccharides that are easier for termites to assimilate. From our BLASTx hits, we could determine the various species present in our dataset and based on previous work, we could infer the role played by the termite symbionts in cellulose digestion, nitrogen fixation and recycling, vitamin production, acetogenesis for energy production among others (see Table 2.3).

Majority of the reads in the dataset had little or no similarity to existing databases. This highlights the potential for discovery of new species or the pres-

Table 2.2. Summary of results on Acid Mine Drainage and simulated dataset

Reads	Termite Metagenome	Acid Mine Drainage	Simulated Dataset
Sequencing Technology	Sanger	Shotgun	454
Number of reads	131637	40000	32530
Average length in bps (median)	540 (594)	892 (790)	105 (106)
GC content		47%	39%
Reads with BLASTx hits	73%	96%	80 %
# genera/#species	2935/6727	1050/4935	2941/11991
Reads with significant identity	47%	89%	76 %
# genera/#species	300/2598	52/100	19/100
# taxon-based clusters	59%	100%	82%
# composition-based clusters	41%	0 %	18 %

Table 2.3. Functions of symbionts in Termite Metagenome

Kingdom	Functions and Characteristics	Species
Protists	Cellulose Digestion	Parabasalia
	Xylan degradation	Trichomonadida
	Anaerobic	Hypotrichomonadida
	Occur in Mitochondria	Spirotrichonymphida
Archaea	Methanogenesis	Methanobrevibacter
Bacteria	Nitrogen fixation	Actinomycetales, Bacillales
	Sulfate Reduction	Desulfovibrio
	Fermentative and Acidogenic	Lactobacter, Enterobacter
	Acetogenesis	Spirochetes (Treponema)

ence of non-protein coding genes. Earlier, we had about 47% of reads grouped into taxon-based clusters. After performing this step, we had an additional 12% of the reads that could be grouped into the existing taxon-based clusters. The remaining reads were grouped into similarity-based clusters. Using the BLAST, Cluster and then Assemble approach, we are able to taxonomically characterize the dataset at read level, identify the species present and quantify the abundance of each species. Table 2.2 summarizes the results of SimComp on Termite gut metagenome, Acid Mine Drainage (AMD) and the simulated dataset described in this chapter.

2.6 Conclusion

In this chapter, we proposed SimComp, a soft clustering method that allows complete and accurate characterization of short metagenome reads that come from a spectrum of known and unknown species. We clustered a simulated dataset using a hybrid of comparative and composition based method. The overlap between the clusters accommodates the ambiguity associated with metagenomic data. It does not require assembled contigs or training on a reference set, nor does it make any assumptions on the number of species or the nature of the dataset.

The oligomer composition of reads as short as 100 bps does not provide sufficient signal to differentiate between species. For best results, we would like to test our algorithm on metagenome datasets with larger read length. Phenomena such as polymorphism and horizontal gene transfer can complicate phylogenetic clustering. As proposed in this chapter, the soft boundary between clusters has the ability to capture such misplacements providing interesting insights into the data. We believe soft clustering has a promising role in classifying metagenome reads and we wish to investigate its scope in the future.

Chapter 3

A Naive Bayes Mixture Model

In this chapter, we formulate an unsupervised naive Bayes multi-species, multi-dimensional mixture model for reads from a metagenome. We use the proposed model to cluster metagenomic reads by their species of origin and to characterize the abundance of each species. We model the distribution of word counts along a genome as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare. We employ either a mixture of Gaussians or mixture of Poissons to model reads within each bin. An additional reason to use these distributions is their flexibility and ease of parameter estimation. Such a paradigm characterizes the compositional heterogeneity of the words along a genome, signifying its genome signature. Further, we handle the high-dimensionality and sparsity associated with the data, by grouping the set of words comprising the reads, resulting in a two-way mixture model. Finally, we derive an unsupervised Expectation Maximization algorithm for the models. Our method provides a general statistical framework for modeling metagenome reads. We demonstrate the accuracy and applicability of this method on simulated and real metagenomes. Our method can accurately cluster reads as short as 100 bps and estimate the species abundance as well. Our method outperforms LikelyBin, another unsupervised composition-based binning method for metagenomes, on datasets of varying abundances, divergences and read lengths.

3.1 Background and Motivation

One of the most common genome signatures is the frequency of occurrence of words (or oligomers) in a DNA sequence[31]. In our method, we model each cluster, containing reads from a species, as a function of probability distributions of words comprising them. The inherent basis of this method is that the set of reads sequenced from a species have a characteristic genome signature that distinguishes it from reads belonging to other species. The distribution of word counts along a genome can be approximated as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare[53]. We propose an unsupervised multi-dimensional Naive Bayes Poisson mixture model and derive an Expectation Maximization algorithm for the same. The corresponding algorithm for Gaussian mixture model can be derived similarly. At times, longer words tend to be more discriminatory than the shorter ones[63]. However, with the increase in the length of words, the dimensionality of the data increases exponentially, while the word counts become sparse. To tackle high dimensionality and sparsity of word counts, we impose a clustering structure on the word counts as well. Such a model is called a two-way mixture model. In essence, the proposed method provides a general statistical framework for associating each read with its species of origin, based on its genome signatures.

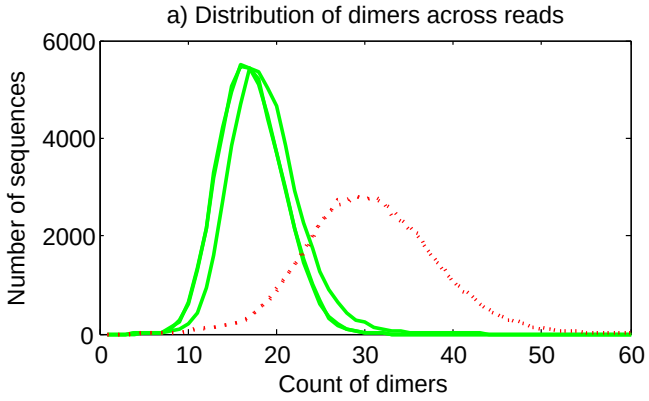


Figure 3.1. Distribution of dimers and pentamers across 50,000 reads sampled from the genome of Haemophilus Influenzae(Only a few distributions are shown). Distribution of dimers tends to Gaussian, two groups can be observed

A genome signature is a compositional parameter reflecting the relative abun-

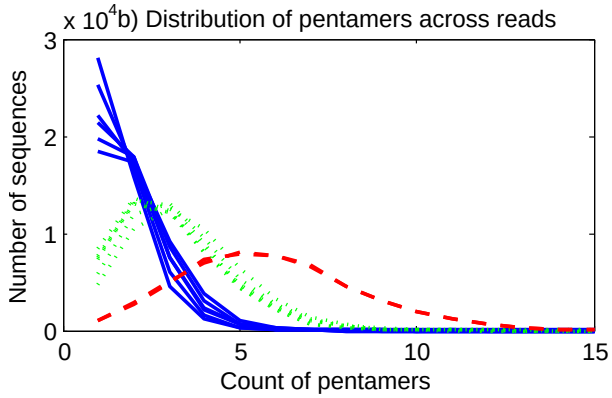


Figure 3.2. Distribution of dimers and pentamers across 50,000 reads sampled from the genome of Haemophilus Influenzae(Only a few distributions are shown). Distribution of pentamers tends to Poisson, three groups are seen.

dance of different words along the genome. In general, it is similar between closely related species and dissimilar between non-related species. Some words that are deemed to be biologically significant, are very common in a genome, while others may never be encountered[8]. Composition-based methods use genome signatures to ascertain the origin of the DNA reads. The underlying basis is that the distribution of words in a DNA is specific to each species and undergoes only slight variations along the genome. By establishing the dictionary of words used by a species and their frequency of occurrence, one can point out the basic words of the genome[19].

Literature abounds in methods that study the statistical distribution of the word locations along a sequence and word frequencies[53, 55]. The exact distribution of count of words is known under the hypothesis that the letters are independent (Bernoulli) or under the Markov model. However, in practice, it is extremely time consuming to compute the exact distribution for long sequences or for frequent words. Hence, two kinds of approximations exist. Distribution of word counts along a genome can be approximately modeled as a Gaussian distribution for short words (that are more frequent), or a Poisson distribution for longer words (that are rare)[53].

A metagenomic dataset consists of reads from different species. The reads sampled along a genome of a species will reflect its genome signature. As different words occur with different frequencies along the genome, each word follows its

own distribution. Thus, reads belonging to a species can be modeled as a multi-dimensional distributions of words (one dimension for each word) comprising them. Figure 3.1 and 3.2 illustrates the distribution of dimer and pentamer counts across reads sampled from the genome of Haemophilus Influenzae (for the purpose of clarity, only a few distributions are shown). We see that count of each dimer (a short word) across the reads, tends to a Gaussian distribution with a different mean and standard deviation and that of a pentamer tends to a Poisson distribution. Hence, the problem of clustering metagenomic reads can be cast as a multi-dimensional mixture of Gaussians (or Poissons for longer words) where distribution of each word is modeled as a Gaussian (or Poisson). In other words, this corresponds to the multi-dimensional Naive Bayes model, where each dimension is modeled as a uni-modal Gaussian (or Poisson) distribution. Such a general statistical model takes into account the compositional heterogeneity of words along the genome.

3.1.1 Multi-species Multi-dimensional Mixture of Distributions

In this chapter, we formulate an unsupervised multi-dimensional Poisson mixture model for clustering reads within a metagenome by their species of origin. We propose to model the reads from a species as a multi-dimensional distribution of the words comprising them. Therefore, each cluster is represented by the distribution of word counts within the species. The multi-dimensional model for Gaussian mixtures can be derived analogously. We present the results for both the models.

Mixture models cover the data well, i.e. dominant patterns in the data are captured by the component distributions. They allow better approximations of the true distributions and their parameters are relatively easy to estimate[62]. An additional advantage of using generative models is that they are flexible and can handle a large number of classes. For instance, a mixture of Poissons can be multi-modal, while a Poisson distribution is always uni-modal.

We begin with a metagenome, $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$, containing N reads from M species. Let α_m be the proportion of species m in the dataset, with $\sum_{m=1}^{M} \alpha_m =$

1. We assume that $\mathbf{X}$ is observed and is governed by some density function $p(\mathbf{X}|\Theta)$ with parameter Θ. Our goal is to cluster the reads by their species of origin, based on the frequency of words that appear in the reads. For every species m, we want to determine α_m, its proportion in the dataset, and Θ, the parameter governing the distribution of words within the reads. Let $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$, be the cluster labels. We assume that $y_i = m$ for $m \in 1, ...M$, if the i^{th} read belongs to the m^{th} species. Also, $p(y_i = m) = \alpha_m$. Cluster label $\mathbf{Y}$ is unknown. We call $(\mathbf{X}, \mathbf{Y})$, the complete dataset.

For a word of length l, we obtain $p = 4^l$ different words (combinations of A, C, T, G), denoted by $W = \{w_1, w_2, ..., w_p\}$. Each read $\mathbf{x_i}$ is represented by a p-dimensional feature vector, $\mathbf{x_i} = \{x_{i1}, x_{i2}, ..., x_{ip}\}$, where x_{ij} is the count of word w_j in read $\mathbf{x_i}$. We model the distribution of words within every species m by a multi-dimensional Poisson distribution, say $\boldsymbol{\lambda_m} = \{\lambda_{m1}, \lambda_{m2}, ..., \lambda_{mp}\}$. That is, given that read $\mathbf{x_i}$ belongs to species m, the distribution of each word w_j is Poisson with parameter λ_{mj}, where $m = 1, 2, ..., M$ and $j = 1, 2, ...p$.

$$p(w_j|\lambda_{mj}) = \phi(w_j|\lambda_{mj}) = \frac{e^{-\lambda_{mj}}\lambda_{mj}^{x_{ij}}}{x_{ij}!} \tag{3.1}$$

We assume independence between features of read vector. The probability of a read $\mathbf{x_i}$, given it belongs to species m is:

$$p(\mathbf{x_i}|y_i = m, \Theta) = p(\mathbf{x_i}|\lambda_m) = \prod_{j=1}^{p} \phi(x_{ij}|\lambda_{mj}) \tag{3.2}$$

At first glance, it might seem imprudent to represent a read as a collection of words comprising it, because it leads to the loss in information about the sequencing read. Strictly speaking, even if the sequence of bases in a DNA are independently and identically distributed, distribution of word occurrences are not independent, due to overlaps[53]. Bayesian networks or belief networks can be used to represent the conditional dependencies between the words comprising the reads[46]. Although, in practice, methods for exact inference in Bayesian networks are often computationally expensive. An attractive alternative to Bayesian networks is the Naive Bayes algorithm that assumes independence between the different features of the read. This assumption makes the otherwise complicated

problem tractable. Naive Bayes is known to perform well on complex models and takes time that is linear in the number of components. In addition, lost information can be restored at later stages. In this chapter, we have presented the formulation of mixture models with the assumption that the different features (word counts) of the read are independent of each other. We outline the Expectation Maximization(EM) algorithm below.

3.1.2 Parameter Estimation

To initialize the estimation algorithm, we randomly assign each read to a cluster m. The posterior probability $q_{i,m}$ is set to 1, if read i is assigned to cluster m and 0 otherwise. With the initial posterior probabilities, a Maximization-step (M-step) is derived to obtain the initial parameters. The EM iterations then follow as below.

Expectation Step: We estimate the posterior probability $q_{i,m}$ of read $\mathbf{x_i}$ belonging to species m. By Bayes theorem, we have,

$$p(y_i = m|\mathbf{x_i}, \Theta) = \frac{\alpha_m.p(\mathbf{x_i}|\lambda_m)}{\sum_{k=1}^{M} \alpha_k.p(\mathbf{x_i}|\lambda_k)} = q_{i,m}$$

$$q_{i,m} \propto \alpha_m. \prod_{j=1}^{p} \phi(x_{ij}|\lambda_{mj}) \text{ subject to } \sum_{m=1}^{M} q_{i,m} = 1 \tag{3.3}$$

Maximization Step: The M-step uses $q_{i,m}$ to compute the expectation of complete data log likelihood,

$$\begin{aligned} Q(\Theta^{(t+1)}, \Theta^{(t)}) &= E_{p(Y|X,\Theta)}[\log p(\mathbf{X}, \mathbf{Y}|\Theta)] \\ &= \sum_{m=1}^{M} \sum_{i=1}^{N} p(y_i = m|\mathbf{x_i}, \Theta^{(t)}). \log(p(\mathbf{x_i}, y_i = m|\Theta^{(t+1)})) \\ &= \sum_{m=1}^{M} \sum_{i=1}^{N} \left(q_{i,m}. \log(\alpha_m.p(\mathbf{x_i}|\lambda_m))\right) \end{aligned}$$

We also take into account the constraint, which requires that α_m's sum to 1 by adding a Langrange multiplier.

$$Q(\Theta^{(t+1)}, \Theta^{(t)}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \left(q_{i,m}. \log(\alpha_m . p(\mathbf{x_i}|\lambda_m)) \right) + \beta(\sum_{m=1}^{M} \alpha_m - 1)$$

We maximize the above expression with respect to the parameters, $\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta^{(t+1)}, \Theta^{(t)})$ and update the parameters,

$$\alpha_m^{(t+1)} = \frac{\sum_{i=1}^{N} q_{i,m}}{N} \ , \ \lambda_{mj}^{(t+1)} = \frac{\sum_{i=1}^{N} q_{i,m}.x_{ij}}{\sum_{i=1}^{N} q_{i,m}} \tag{3.4}$$

Finally, these two steps are repeated as necessary. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function.

3.1.3 Word Grouping

Higher order words are known to be more discriminative than shorter ones[63]. With the increase in length of the word, there are two major consequences that need to be addressed. Firstly, the distribution of words tends to Poisson and not Gaussian (by law of rare numbers), see Figure 3.1 and 3.2. Secondly, the length of the read vector grows exponentially (e.g, for $l = 10, 4^l \approx 10^6$). With increase in dimensions, many words will tend to have similar distributions and hence, can be clustered together into a "word group". At the same time, the number of distinct words in any read is usually substantially smaller than the number of dimensions. That is, the feature matrix becomes high-dimensional and sparse. Hence, the model may fail to predict the true feature distribution of different components. Therefore, dimension reduction becomes necessary before estimating the components in the model. However, reduction of the number of words using feature selection cannot be too aggressive, otherwise the clustering accuracy will suffer.

In this chapter, we handle the above challenge by "word grouping". A supervised two-way Poisson Mixture Model with word grouping was originally proposed by Li *et al.* for simultaneous document classification[36]. Such a two-way clus-

tering involves simultaneous clustering of reads as well as of words. The clusters means are regularized by dividing the words into groups and constraining the parameters for the words within the same group to be identical. The grouping of the words is not pre-determined, but optimized as part of the model estimation. This implies that for every group, only one statistic for all the words in this group is needed to cluster reads. For instance, in Figure 3.2, we observe the distribution of pentamers falls into three distinct group. Therefore, words following similar distributions can be clustered together into a "word group".

We extend our formulation to an equivalent two-way unsupervised Poisson mixture model in order to simultaneously cluster word features and classify reads and derive an Expectation Maximization algorithm to estimate its parameters. Figure 3.3 depicts the paradigm for two-way mixture model of reads. Note that we make a distinction on the use of "cluster" to refer to binning of reads belonging to the same species and "group" to refer to binning of words within read in a cluster.

Recall that the genome signature is similar between closely related species and dissimilar between non-related species. The parameter constraint implies that words have the same distribution within each cluster. Therefore, we can assume that within each cluster, words in different reads have equal Poisson parameters, while for reads in different clusters, words may follow different Poisson distributions. For simplicity, we assume that all clusters have the same number of word groups. It is trivial to extend to the case where different clusters may have different number of word groups[50].

Let $l \in 1, ..., L$ denote the word groups. We define a group assignment function $c(m, j) \in 1, 2, ..., L$, which denotes the group to which word w_j belongs in class m. Words in the same word group will have the identical parameters, i.e. $\lambda_{mk} = \lambda_{mj} = \theta_{m,l}$, if $c(m, k) = c(m, j)$. The group assignments of the words vary from cluster to cluster. Let the number of words in group l of class m be η_{ml}. The likelihood of $\mathbf{x_i}$ is now:

$$p(\mathbf{x_i}|\lambda_m) = \prod_{j=1}^{p} p(x_{ij}|\lambda_{mj}) = \prod_{j=1}^{p} p(x_{ij}|\theta_{m,c(m,j)}) \tag{3.5}$$

Now, we can perform clustering using no more than ML dimensions. Word

grouping leads to dimension reduction in this precise sense.

We can derive an EM algorithm similar to the one outlined above to estimate the Poisson parameters $\theta_{m,l}$ where $m \in 1, ..., M, l \in 1, ..., L$, the group assignment function $c(m, j) \in 1, ..., L$, where $m \in 1, ..., M, j \in 1, ..., p$ and the prior mixture components α_m, for $m \in 1, ..., M$. We initialize by setting each value of the group assignment function $c(m, j)$ randomly to a number in $1, ..., L$. We start with the same word group partition for all the clusters, i.e. $c(m, j)$'s are initially identical over m. We update the parameters as given below:

$$\alpha_m^{(t+1)} = \frac{\sum_{i=1}^{N} q_{i,m}}{N} \ , \ \theta_{m,l}^{(t+1)} = \frac{\sum_{i=1}^{N} q_{i,m} \cdot \sum_{j \in l} x_{ij}}{\eta_{ml} \sum_{i=1}^{N} q_{i,m}} \text{ where } c(m, j) = l \tag{3.6}$$

Once $\theta_{ml}^{(t+1)}$ is fixed, the word cluster index $c^{(t+1))}(m, j)$ can be found by doing a linear search over all components:

$$c^{(t+1)}(m, j) = arg \max_{l} \sum_{i=1} q_{i,m}(x_{ij} \log \theta_{m,l}^{(t+1)} - \theta_{m,l}^{(t+1)}) \tag{3.7}$$

3.1.4 Naive Bayes Mixture of Multinomials

If $(X_1, X_2, .., X_p)$ are independent Poisson variables with parameters, $\lambda_1, \lambda_2, .., \lambda_p$ respectively, then the conditional distribution of $(X_1, X_2, .., X_p)$ given that $X_1 + X_2 + ... + X_p = n$ is multinomial with parameters λ_j/λ, where $\lambda = \sum \lambda_j$, i.e. $Mult(n, \pi)$, where $\pi = (\lambda_1/\lambda, \lambda_2/\lambda, ..., \lambda_p/\lambda)$[21].

The above theorem implies that the unconditional distribution $(X_1, X_2, ..., X_p)$ can be factored into a product of two distributions: a Poisson for the overall total, and a multinomial distribution of X, $X \sim Mult(n, \pi)$. Therefore, the likelihood based inferences about π are the same whether we regard $X_1, X_2, .., X_p$ as sampled from p independent Poissons or from a single multinomial. Here, n refers to the length of the reads and our interest lies in the proportion of words in the reads. Any estimates, tests, inferences about the proportions will be the same whether we regard n as random or fixed.

We can now derive the Naive Bayes mixture of Multinomials as standardized mixture of Poissons. We assume that the distribution of words within the reads of a species is governed by the parameters of a multinomial distribution

$\mathbf{\Theta} = (\theta_\mathbf{1}, \theta_\mathbf{2}, ..., \theta_\mathbf{m})$, where each $\theta_\mathbf{m}$ is the parameter for species m and is given by $\theta_\mathbf{m} = (\theta_{m1}, \theta_{m2}, .., \theta_{mp})$. Therefore, the likelihood of the data will be,

$$P(\mathbf{x_i}|y_i = m) = P(\mathbf{x_i}|\theta_\mathbf{m}) = \frac{n_i!}{\prod_{j=1}^{p} x_{ij}!} \prod_{j=1}^{p} \theta_{mj}^{x_{ij}} \tag{3.8}$$

The sum of the probabilities satisfies the constraint $\sum_{j=1}^{p} \theta_{mj} = 1$. The EM algorithm for Naive Bayes mixture of Multinomials can be derived similarly and we only give the final set of equations.

$$\alpha_m = \frac{\sum_{i=1}^{N} q_{i,m}}{N} \ \& \ \theta_{mj} = \frac{\sum_{i=1}^{N} q_{i,m}.x_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{p} q_{i,m} x_{ij}} = \frac{\sum_{i=1}^{N} q_{i,m}.x_{ij}}{\sum_{i=1}^{N} q_{i,m} n_i}$$

If we assume the length of each read to be a constant n, we get the same results as that with Poisson distribution, hence the two distributions are equivalent in modeling the distribution of words within reads of a species. Also, since the Multinomial distribution is single distribution, we do not perform a two-way dimension reduction on it.

3.2 Results

3.2.1 Simulated metagenomes

The algorithm has been implemented in Matlab and C. The space and time complexity scale linearly with the number of reads and species. Space complexity scales quadratically with the number of dimensions in the search space. Our method converged for all the cases we tested and was robust to the choice of initial conditions(Figure 3.4).

Metagenomics being a relatively new field, lacks standard datasets for the purpose of testing clustering algorithms[41]. As the "true solution" for sequence data generated from most metagenomic studies is still unknown, we focused on synthetic datasets for benchmarking. We also apply our method to the actual Acid Mine Drainage dataset to identify the dominant species. In order to test the accuracy of our proposed method, we used Metasim to simulate synthetic metagenomes[27]. Metasim takes as input the sequencing technology to

be used (Sanger, 454, Exact), a set of known genomes, length of the reads and an abundance/coverage profile which determines the relative abundance of each genome in the simulated dataset. The genomes used for generating the synthetic metagenomes were downloaded from National Center of Biotechnology Information (NCBI). We generated datasets with reads of lengths between 50 and 1000 bp and various abundance ratios. In the first part of this section, we demonstrate the performance of the multi-dimensional Gaussian mixture model on several datasets. A default word length of 2 is used. Additionally, as the number of dimensions is relatively small, we do not perform word grouping. Next, we describe the results using the two-way Poisson mixture model with a word length of 5. The method has been implemented for word lengths from 2 to 9. In order to calculate the clustering accuracy, we assign each cluster to the source species that is most frequent in the cluster. Accuracy is given by the percentage of total correct read assignments.

The number of species in each dataset is supplied as an input. Determining the number of clusters from a statistical perspective is a difficult problem and has been addressed by [65]. Previously, 16s/18s rDNA have been used for phylotyping and assessing species diversity using a rarefraction curve[13]. Tools such as MetaPhyler and TreePhyler can be used for making an educated guess of the number of species[39, 60]. Estimating species diversity is still an active area of research and we do not address it in this chapter.

Experiments in the 1960s and 1970s have shown that the dinucleotide relative abundance in a genome is a remarkably stable property[30, 57]. Closely related organisms display more similar dinucleotide composition than do distant organisms[31]. In [9], the authors proposed a measure of intergenomic difference between two sequences f and g, called the average dinucleotide relative abundance,

$$\delta^*(f,g) = \frac{1}{16}\sum_{X,Y} |\rho^*_{XY}(f) - \rho^*_{XY}(g)| \tag{3.9}$$

where $\rho^*_{XY}(f) = \frac{f^*_{XY}}{f^*_Y f^*_Y}$ and f^*_X denotes the frequency of X in f. A measure of intergenomic difference was obtained by comparing different genome signatures. In order to assess the robustness of our method, we test it across datasets represen-

tative of δ^* values ranging from 34 to 340. In general, lower δ^* values correspond to "closely related species" and higher values correspond to "distant species".

In Figure 3.5, we plot the performance of our proposed multi-dimensional Poisson model over 450 datasets with δ^* values ranging from 34 to 340. We observed a positive correlation between the intergenomic difference and the accuracy of our method, as also noted in [33]. The initial increase in the accuracy with word length is justified by the increased discriminative power of higher order words. However, any further increase in word length has to be accompanied by dimension reduction, otherwise owing to the high dimensional and sparse nature of feature matrix, the accuracy begins to drop.

In Figure 3.6, we compare the accuracy of our proposed multi-dimensional Gaussian model with two other unsupervised composition-based methods LikelyBin[33] and Scimm[32] on several datasets. Default parameters are used for these algorithms. We varied the read length between 200 to 500 bp, δ^* values from 60 to 300 and the abundance ratio up to 1:5. Note that the distribution of dimers tends to a Gaussian. As the number of dimensions is relatively small ($4^2 = 16$), the algorithm performs well without word grouping. Our method clearly outperforms LikelyBin and performs as well or better than Scimm on most instances. Another point worth noting from the figure is that our method's error rate is bounded by 10% for datasets with read length as short as 200 bp.

We analyzed the accuracy and applicability of our method on binning reads from low complexity communities, containing 3-5 species (see Table 3.1). With the increase in number of species, there was a slight degradation in performance, though the accuracy was consistently above 85%. This is in agreement with the results from the 2 species dataset, considering that the total coverage of each species is much lower in a multi-species dataset (Reads from *B. Burgdorferi* form only 6% of the 5^{th} dataset).

Next, we evaluated the robustness of our method to changes in the abundance ratio between species as well as the length of the reads. We simulated three sets of metagenomes with two species each, at different abundance ratios. We varied the abundance ratio from 10:1 to 1:10 in stages, for the two species. From Figure 3.7,

Table 3.1. Performance of Gaussian Mixture Model (without word grouping) on datasets containing more than 2 species, at various abundances on reads of length 500 bp. AR stands for Abundance Ratio

Species	AR	#reads	Accuracy(%)
T. Thermophilis	1	50000	87.51
A. Vinelandii	3		
N. Meningitidis	2		
E. Coli 536	1	50000	97.01
S. Acidocaldarius	2		
H. Salinarium R1	2		
C. Jejuni RM1221	3	60000	96.61
H. Salinarium R1	2		
E. Coli	1		
P. Horikoshii OT3	3		
S. Erythraea	1	60000	90.28
M. Thermoautotrophicum	1		
B. Burgdorferi ZS7	1		
E. Coli 536	1		
B. Burgdorferi ZS7	1	75000	85.04
C. Jejuni RM1221	1		
E. Coli 536	1		
H. Salinarum R1	1		
P. Horikoshii OT3	1		

we note that there was only a slight drop in performance for extreme abundance ratios. Therefore, the proposed method is suited for binning relatively rare species as well. It is noteworthy to point out that estimates are good at all abundances. In order to test the usefulness of the method for analyzing data produced by the current NGS technologies (especially Solexa and SOLiD) that generate short reads, we tested three datasets of varying δ^* values for read lengths between 50 to 1000 bp. With the decrease in read length from 1000 to 50 bp, the drop in accuracy of our method is bounded by 15 %.

Recall that with the increase in the length of the words and the simultaneous increase in the number of dimensions, the distribution of the words tends to a Poisson and word grouping becomes necessary. In this section, we present the clustering results obtained by estimating the two-way Poisson mixture model with different number of word groups L. We observed the variation in classification accuracy to be more prominent for lower values of L. Therefore, in Table 3.2, we

Table 3.2. Performance of Poisson mixture model on datasets for different values of L and word length of 5. Here, N.W.G stands for no word grouping. The maximum accuracy achieved is in bold. Each dataset contains 50,000 reads of length 500 bp.

Species	$L = 5$	$L = 10$	$L = 30$	$L = 50$	N.W.G
B. Anthracis CI chromosome B. Halodurans C-125	90.61	**91.53**	50.31	91.2	50.32
H. pylori 26695 S. pneumoniae 70585	98.6	**98.79**	98. 73	98.71	98.76
B. Subtilis subsp. spizizenii str. L. Lactis subsp.	89.96	90.34	**90.62**	90.53	50.47

report the results for values of $L < 50$ for a 2 species dataset. If word grouping is not performed, then clustering based on mixture model is essentially the Naive Bayes algorithm with each dimension modeled by a Poisson distribution (last column of Table 3.2). From the results, we can infer that word grouping resulted in considerable increase in accuracy compared to the Naive Bayes algorithm. That is, the characteristic vectors are of a much lower dimension with $L \ll p$. Also, a high clustering accuracy can be achieved using no more than ML dimensions, significantly smaller than the original dimension, 1024. Note that it is difficult to know a priori, the exact value of L that yields the best clustering. However, among the values we tested, lower values of L provided a higher accuracy.

Table 3.3. Comparison of performance of Gaussian mixture model (GMM) with 2-way Poisson mixture model (PMM) for datasets with low δ^* values. Each dataset contains 50,000 reads of length 500 bp.

Species	δ^*	GMM	PMM
M. Leprae, P. Putida	74	75.25	85.24
B. Subtilis , L. Lactis	86	86.23	90.62
H. Pylori , S. Pneumoniae	148	53.48	98.76
H. Salinarum, R. Sphaeroides	153	94.63	98.51
M. Jannaschii, S. Aureus	164	50.0	97.75

In Table 3.3, we compare the performance of our 2-way Poisson mixture model with Gaussian mixture model for datasets with low δ^* values. In real situations, it is difficult to know beforehand, the most discerning order of the word to use. However, from our experiments, we can infer that higher-order word-based models, in general, tend to be more discriminatory than those based on lower order words. If it is known a priori that lower order words (of length 2-3) are more dis-

criminatory in the dataset, then we recommend using a Gaussian mixture model. For other datasets, we use a Poisson mixture model.

Table 3.4. Performance of Poisson Mixture Model (without word grouping) on datasets across various taxonomic ranks. Each dataset contains 50,000 reads of length 500 bp. AR stands for Abundance Ratio

Species	AR	Rank	Accuracy(%)
M. Hyopneumoniae, M. Mycoides	3:2	Genus	95.73
M. Avium, M. Leprae	3:4	Genus	94.22
A. Vinelandii, C. Japonicus	1:1	Family	92.81
M. Leprae, S. Erythraea	1:1	Order	95.58
B. Pertussis, N. Gonorrhoeae	1:2	Class	97.52
A. Parvulum, S. Erythraea	5:1	Class	99.64
R. Prowazekii, S. Meliloti	3:1	Class	99.91

Our method's accuracy in classifying reads from the datasets composed of species across various taxonomic ranks is reported in Table 3.4, we used the Poisson mixture model without word grouping. The error rates are bounded by 10% on all datasets. We can infer that the accuracy is mostly correlated to the phylogenetic distances between the species. For example, reads from datasets containing species with taxonomic differences at the level of class were classified with a very high accuracy.

3.2.2 Real metagenome: Acid Mine Drainage Dataset

The ultimate goal of binning methods is to cluster reads in a real metagenome, by their species of origin. Clustering in real situations is error-prone and affects our final estimates of species abundance. Moreover, evaluating clustering methods on real metagenomes can be problematic as the true taxonomic composition of the data is mostly unknown. The accuracy of unsupervised clustering methods decreases with increase in the complexity of metagenomes and for species present at very low abundances. However, the composition of Acid Mine Drainage metagenome has been substantially characterized and we used this dataset to evaluate the performance of our proposed method[68]. The AMD

microbial community is reported to consist of two dominant populations (*Ferroplasma sp. Type II* and *Leptospirillum sp. Group II*) and three other less abundant ones (*Ferroplasma acidarmanus Type I, Leptospirillum sp. Group III* and *Thermoplasmatales archaeon GpI*). We downloaded the reads, as well as the scaffolds assembled from the reads for the 5 species of the actual AMD dataset, from NCBI. Only 58% of the AMD reads can be mapped back to the assembled scaffolds using BLAST[72]. Therefore, in order to compute the accuracy of our method, we simulated a metagenome with reads sampled from the downloaded scaffolds. The simulated AMD dataset consisted of 110,000 reads of average length 732 bp (average read length in the actual AMD dataset) from the 5 species, in the ratio 4:4:1:1:1. We characterized the dataset in two stages. Notice that the dataset contains reads with two distinct abundance levels. Therefore, we can simplify the problem by first separating the reads into two bins based on their abundance. In the first stage, the reads were grouped into two bins, using Abundance Bin, with a resulting accuracy of 93.3%. The bins corresponding to the abundance levels of 4 and 1 had a cluster purity of 93.2% and 98.2% respectively. In the next stage, we used the reads from each of the bins output by Abundance Bin, as an input to our proposed 2-way Poisson mixture model, to further classify the reads by their species of origin. We used a word length of 5. Our method clustered the reads from the bin containing dominant species into two clusters corresponding to *Ferroplasma sp. Type II* and *Leptospirillum sp. Group II*, with an accuracy of 96.88% (with $L = 10$). The other bin consisted of very few reads from the remaining three species *Ferroplasma acidarmanus Type I, Leptospirillum sp. Group III* and *Thermoplasmatales archaeon GpI*. Our method clustered the reads from this bin into three clusters, with an accuracy of 70.34% (with $L = 10$). This decrease in accuracy can be attributed to the low bin count (see Figure 3.6).

3.3 Discussion

In this chapter, we formulated an unsupervised two-way multi-species, multi-dimensional mixture model to represent reads from a metagenome. We used the proposed model to cluster metagenomic reads by their species of origin and to

characterize the abundance of each species. The distribution of word counts along a genome can be approximated as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare. Therefore, we use a multi-dimensional mixture of Gaussians or Poissons to model the reads from each bin. An additional reason to use these distributions is their flexibility, stability and ease of parameter estimation. Our method is an unsupervised method that does not require any training data. This is critical for success as most metagenomic datasets contain reads from unexplored phyla which cannot be labeled into one of the existing classes. Our probabilistic approach can be used to identify reads which belong to more than one species and occlude the cluster boundaries. Such reads should be further investigated to identify the presence of conserved regions.

Note that our proposed method is primarily a composition-based method that seeks to distinguish between genomes based on their characteristic DNA compositional pattern. Therefore, it cannot distinguish between genomes unless their DNA compositions are sufficiently divergent (see Figure 3.6, dataset with *B. Burgdorferi, C. Jejuni*). It is unlikely that our method will be able to accurately distinguish between strains of the same species. For such datasets, genome signature alone is insufficient for inferring taxonomic relationships reliably. Composition-based methods must be used in conjunction with other similarity-based methods and abundance-based methods to yield better performance.

Note that the two-way Poisson mixture model was originally proposed for classification of documents. In this work, we demonstrate the relevance and applicability of such a general statistical framework for modeling metagenome reads. We have illustrated that the proposed method can accurately classify reads from low to medium complexity datasets into taxon-specific bins, based on genome signatures.

Our framework complements the existing similarity-based and abundance-based methods and hence, can be combined with such methods to obtain a better performance. We intend to develop such hybrid methods in the future that can tackle the problem of classifying sequences in complex metagenomic communities.

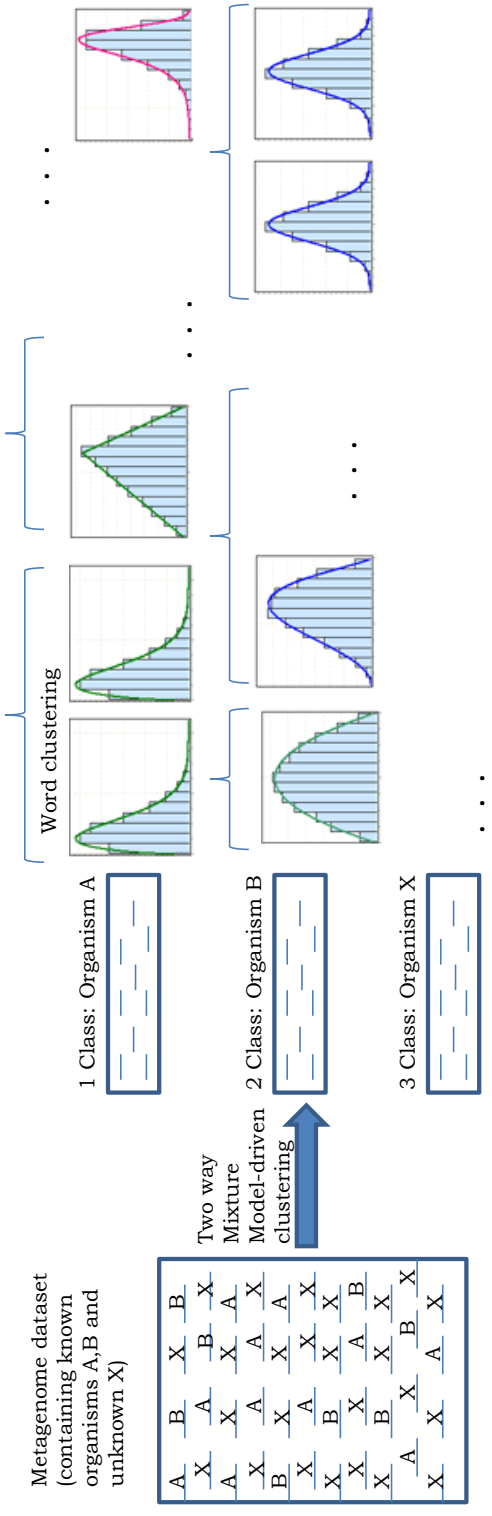


Figure 3.3. Illustration of a Two-way Poisson Mixture Model for Metagenomic Data. Each cluster represents a species and is modeled as a distribution of words comprising it. Each word follows a different distribution. However, not all words in a class have significantly different parameters. Therefore, the words can be divided into groups and words within the same group can be constrained to have identical parameters.

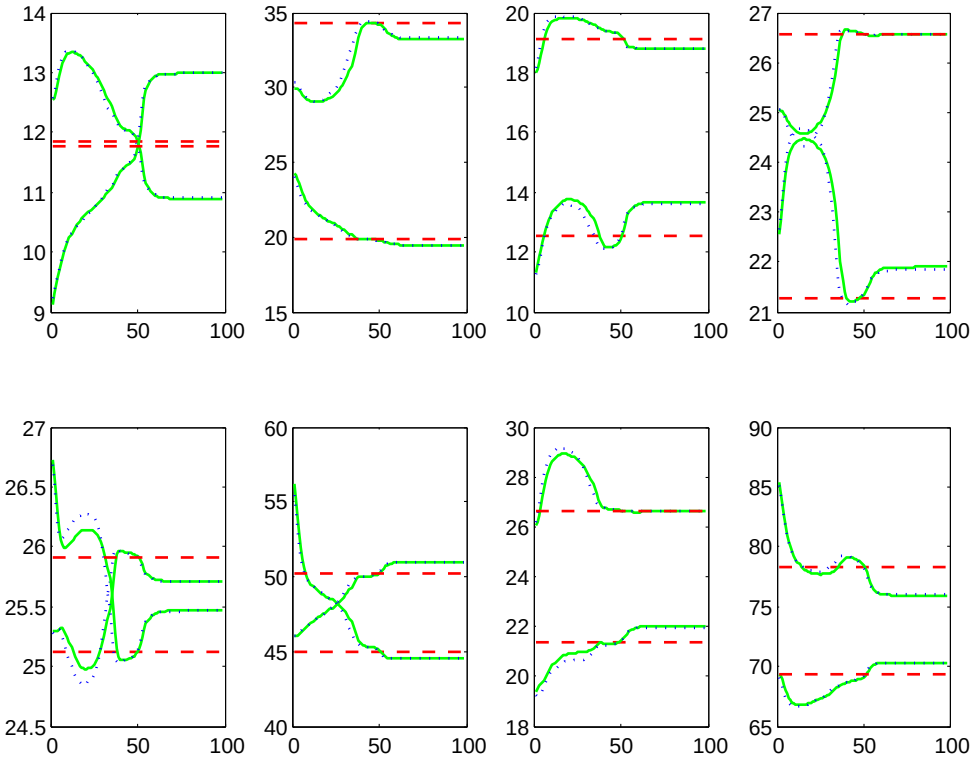


Figure 3.4. Our method converges for all cases tested and is robust to the choice of initial conditions.

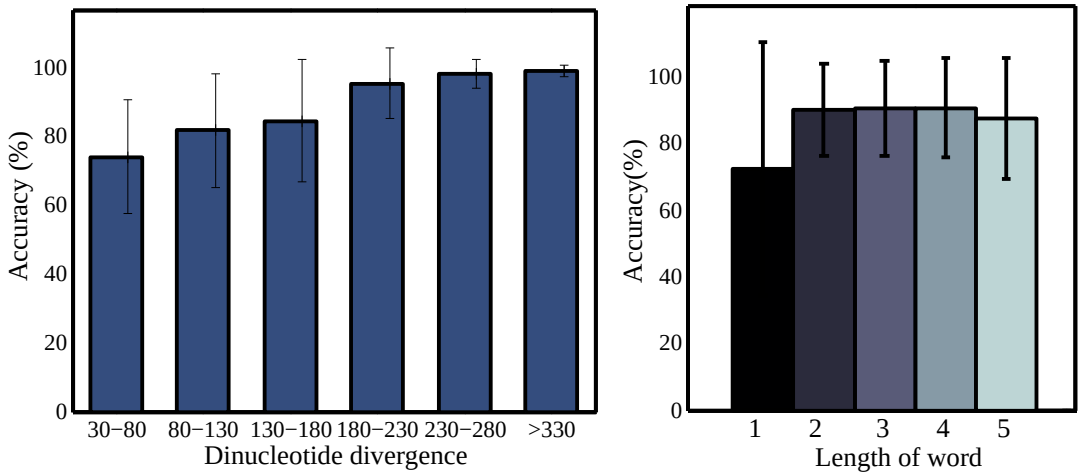


Figure 3.5. Performance of Poisson Mixture Model vs. **Left.** Dinucleotide divergence. **Right.** Word length over 450 datasets with δ^* values ranging from 34 to 340

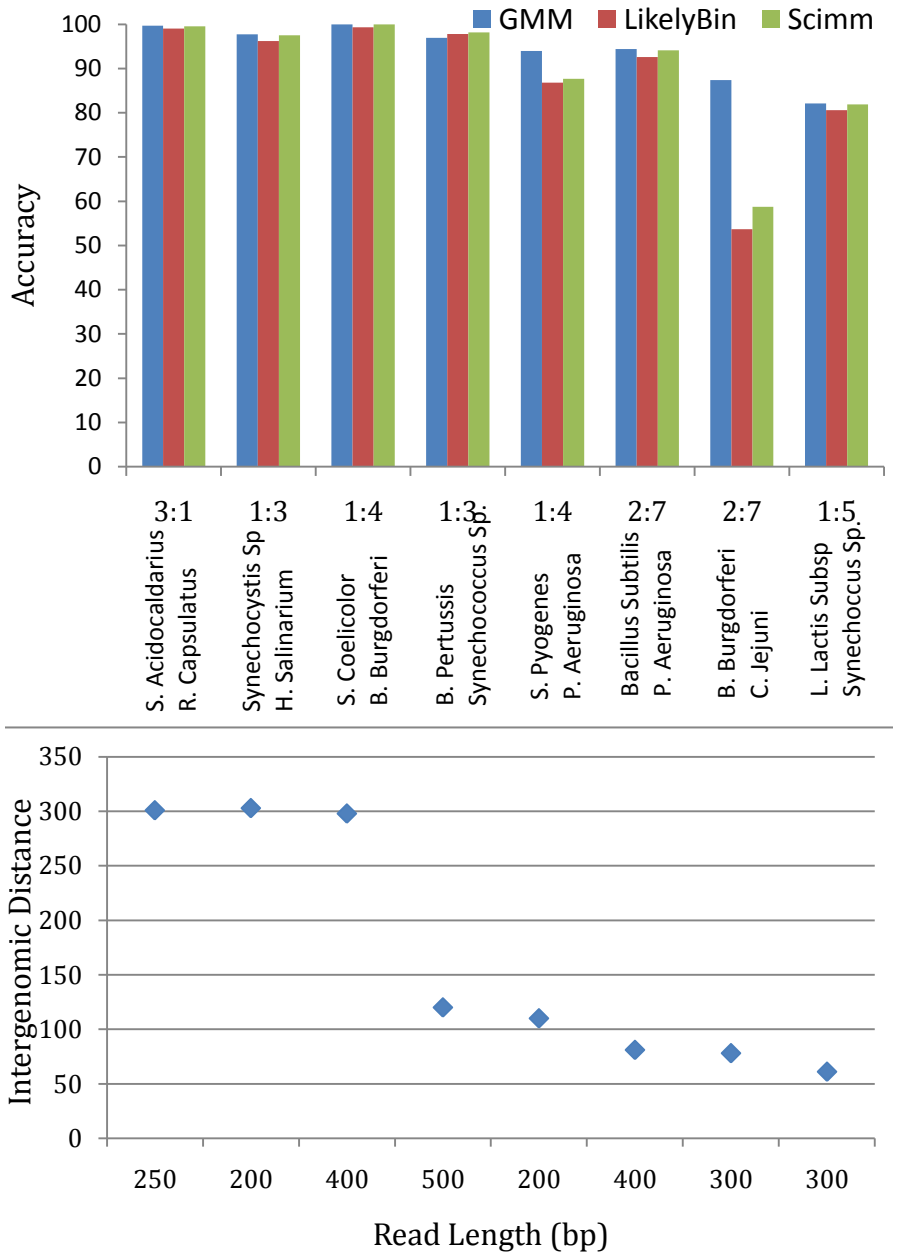


Figure 3.6. GMM stands for Gaussian mixture model (without word grouping). The top figure compares the performance of the three methods on 8 datasets (X-axis shows the abundance ratio and the species contained in the dataset.). The bottom figure plots the δ^* values for the corresponding datasets. The X-axis shows the corresponding read lengths. Here, the δ^* (measured on 50 kb contigs) ranges from 34 to 340.

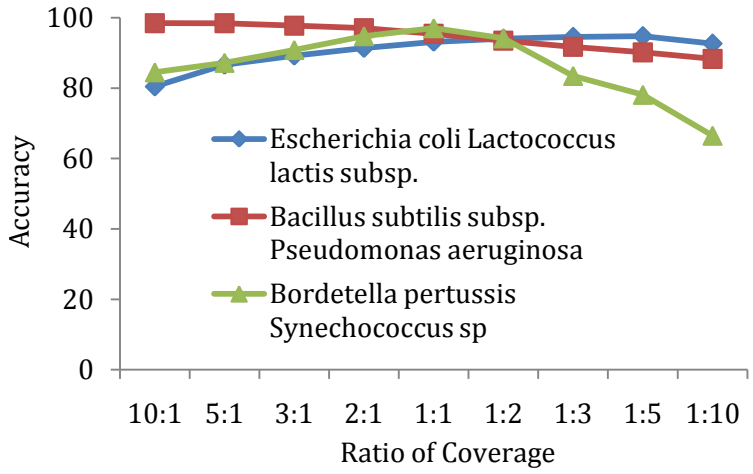


Figure 3.7. Performance of Poisson Mixture Model at different reads lengths (50-1000 bp). Datasets with low δ^* values (100-150) were chosen.

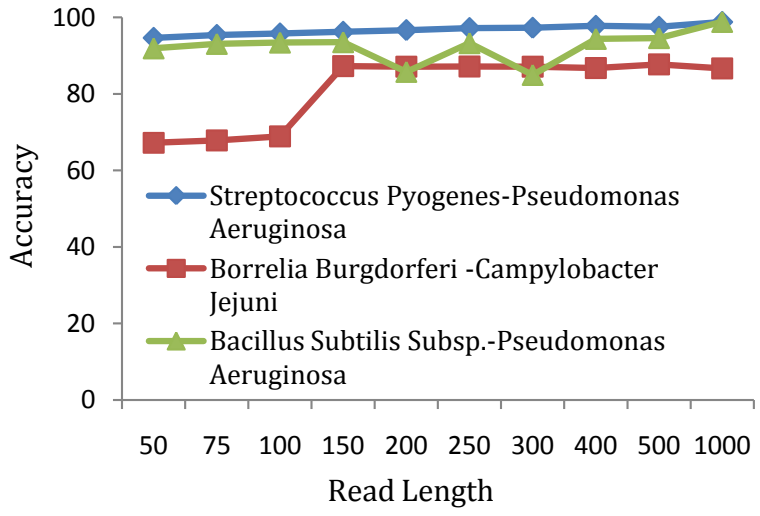


Figure 3.8. Performance of Poisson Mixture Model at different coverage ratios in a 2-species datasets. Datasets with low δ^* values (100-150) were chosen.

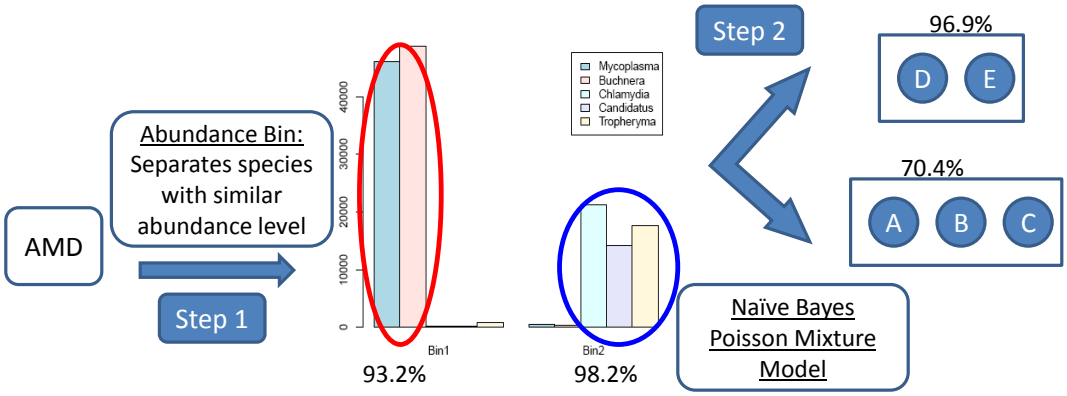


Figure 3.9. Results on the Acid Mine Drainage Dataset

Chapter 4

A Bayesian Mixture Model

In this chapter, we present a new and efficient Bayesian mixture model based on Poisson and Multinomial distributions for clustering metagenomic reads by their species of origin. We use the relative abundance of different words along a genome to distinguish reads from different species. The distribution of word counts within a genome is accurately represented by a Poisson distribution. The Multinomial mixture model is derived as a standardized Poisson mixture model. The Bayesian network efficiently encodes the conditional dependencies between word counts in a DNA due to overlaps and hence is most consistent with the data. We present a two-way mixture model that captures the high dimensionality and sparsity associated with the data. Our method can cluster reads as short as 50 bps with accuracy over 80%. The Bayesian mixture models clearly outperform their Naive Bayes counterparts on datasets of varying abundances, divergences and read lengths. Our method attains comparable accuracy to that of state-of-art Scimm and converges at least 5 times faster than Scimm for all the cases tested. The reduced time taken, by our method, to obtain accurate results is highly significant and justifies the use of our proposed method to evaluate large metagenome datasets.

4.1 Background and Motivation

Mixture models have become popular tools for analyzing biological sequence data. The relevance of mixture models comes from the generative modeling approach

to clustering that can handle large number of classes and incorporate domain knowledge with ease. Mixture models cover the data well i.e. dominant patterns in the data are captured by the component distributions.

Most mixture models assume an underlying multivariate normal distribution. However, the distribution of word counts within a genome vary according to a Poisson distribution[53, 55]. The Poisson distribution is adequately approximated by a normal distribution for short, frequent words with high count. However, when the count is low, the Poisson distribution more accurately represents the data and hence provides advantages over other distributions[70]. An additional reason to use these distributions is their flexibility and ease of parameter estimation. Figure 4.1 illustrates the distribution of dimers and pentamers across reads sampled from the genome of Haemophilus Influenzae (for the purpose of clarity, only a few distributions are shown). Hence, the problem of clustering metagenomic reads where distribution of each word varies according to a Poisson distribution can be cast as a multivariate mixture of Poissons.

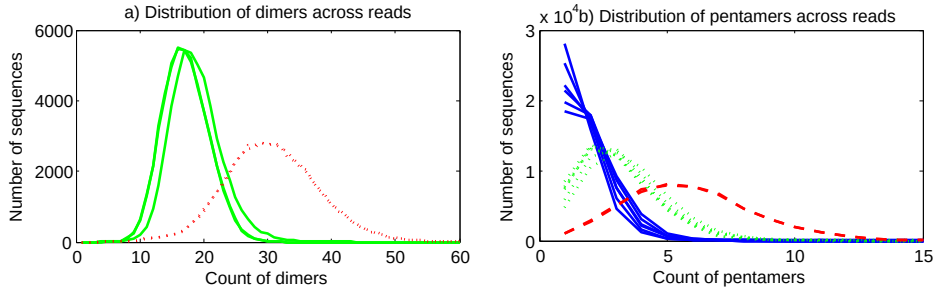


Figure 4.1. This figure illustrates the distribution of dimers and pentamers across 50,000 reads sampled from the genome of Haemophilus Influenzae. a) Distribution of dimers tends to Gaussian and is approximated by a Poisson, two distinct groups can be observed. b) Distribution of pentamers tends to Poisson, three groups can be observed.

Bayesian and Naive Bayes models: Bayesian networks can efficiently represent complex probability distributions. It encodes the joint probability distribution of a set of n variables, $\{X_1, X_2, ..., X_n\}$ as a directed acyclic graph and a set of conditional probability distributions (CPDs). The set of parents of X_i are denoted by Pa_i. Each X_i is conditionally dependent on its parents Pa_i and independent of its non-descendants given its parents. The joint probability

distribution is given by,

$$p(X_1, X_2, ..., X_n|\Theta) = \prod_{i=1}^{n} p(X_i|Pa_i, \Theta) \tag{4.1}$$

Typically, even if the sequence of bases in a DNA are independently and identically distributed, distribution of word counts are not independent due to overlaps and hence, Bayesian networks are ideal for representing the dependencies between words. Though, in practice, methods for exact inference of the structures in Bayesian networks are often computationally expensive. An alternative to Bayesian networks is the Naive Bayes algorithm that assumes independence between the variables. The assumption makes the otherwise complicated problem tractable. Naive Bayes takes time linear in the number of components. Naive Bayes is the simplest Bayesian network that does not represent any variable dependencies. In the last chapter, we presented a Naive Bayes mixture model for clustering metagenome reads. The motivation in this chapter is to overcome the bottleneck of Naive Bayes by taking into account the conditional dependencies between the word counts within the reads. We focus on developing a tractable Bayesian network for a mixture of Poisson and Multinomial distributions. We will use Bayesian networks to specify the structure of the network and to learn the parameters as well.

Willse *et al.* described the Naive Bayes mixture of Poissons and the corresponding standardized mixture of Multinomials in their paper [70] and used it to segment secondary ion mass spectrometry images into chemically homogeneous regions. The authors implicitly assume that the variables within a class are independent. Later, we compare the Bayesian probability models proposed in this paper to their Naive Bayes counterpart to determine the model most consistent with the data and hence can be regarded as an approximate data-generation mechanism. In the next section, we will develop a Bayesian network for a mixture of Poissons and Multinomials. Note that all the methods in this chapter have been explained using the example of clustering reads in a metagenome by their species of origin. However, the proposed methods can be extended to any discrete sequence data.

4.2 Methods

We are given a metagenome dataset, $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$, containing N reads from M species. Let α_m be the proportion of species m in the dataset, with $\sum_{m=1}^{M} \alpha_m = 1$. We assume that $\mathbf{X}$ is observed and is governed by some density function $p(\mathbf{X}|\Theta)$ with parameter Θ. Our goal is to cluster the reads by their species of origin, based on the frequency of words that appear in the reads. For every species m, we want to determine α_m, its proportion in the dataset, and Θ, the parameter governing the distribution of words within the reads. Let $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$, be the cluster labels. We assume that $y_i = m$ for $m \in 1, ...M$, if the i^{th} read belongs to the m^{th} species. Also, $p(y_i = m) = \alpha_m$. Cluster label $\mathbf{Y}$ is unknown. We call $(\mathbf{X}, \mathbf{Y})$, the complete dataset.

We use Bayesian networks to represent the conditional dependencies between words. The distribution of each word is independent of its non-descendants given its parents. Let read $\mathbf{x}$ be of length n , $\mathbf{x} = (c_1 c_2 ... c_n)$, where each $c_k \in (A, C, T, G)$. We assume that the probability of the read is determined by a set of $p = 4^l$ probabilities corresponding to words of length l.

$$p(\mathbf{x}|\Theta) = p(c_1 c_2 ... c_l) \prod_{k=l+1}^{n} p(c_k | c_{k-l} ... c_{k-1}, \Theta) = \prod_{k=1}^{n} p(c_k | pa_k, \Theta) \tag{4.2}$$

Due to word overlaps, we can define conditional probability dependencies. In a read, any given nucleotide c_k can be preceded by its parents pa_k in the read, where $c_j \in (A, C, T, G)$ and $pa_k \in \{pa_k^1, pa_k^2, .., pa_k^p\}$ denote different word configurations of parents. In the next section, we will formulate the Bayesian mixture of Poissons from first principles. In section 4.2.2, we present the two-way Bayesian mixture of Poissons that uses "word grouping" to handle high-dimensionality and sparsity associated with the metagenome. In section 4.2.3 and 4.2.4, we present the Bayesian mixture of Multinomials as standardized Bayesian mixture of Poissons and the corresponding two-way Bayesian mixture of Multinomials respectively.

4.2.1 Bayesian Mixture of Poissons

We represent each read $\mathbf{x_i}$ by a $4Xp$ matrix $\mathbb{N}_i = \{N_i(c_k|pa_j) : j = 1, ..p\}$ and $c_k \in (A, C, T, G)$, where $N_i(c_k|pa_j)$ is the count of the number of occurrences of

parent word pa_j followed by nucleotide c_k in read $\mathbf{x_i}$. The distribution of words within the reads of a species follow the parameters of a Poisson distribution, $\mathbf{\Theta} = (\lambda_\mathbf{1}, \lambda_\mathbf{2}, ..., \lambda_\mathbf{m})$, where $\lambda_\mathbf{m} = ((\lambda_{m,c_k|pa_j})_{\forall c_k})_{\forall pa_j}$, i.e., each local species distribution is a collection of Poisson distributions, one for every configuration pa_j of parents and c_k, and has the same parameters across reads of a species.

$$\begin{aligned}
&\mathbf{\Theta} = \{\lambda_\mathbf{m} : \forall m \in 1, .., M\} \\
&\lambda_\mathbf{m} = \{\lambda_{m,c_k|pa_j} : \forall c_k \in (A, C, T, G) \text{ and } \forall pa_j \in \{pa_j^1, pa_j^2, .., pa_j^p\}\} \quad (4.3)
\end{aligned}$$

Therefore, the likelihood of the data will be,

$$\begin{aligned}
p(\mathbf{x_i}|y_i = m) = p(\mathbf{x_i}|\lambda_\mathbf{m}) &= \prod_{pa_j}\prod_{c_k} p(N_i(c_k|pa_j)|\lambda_{m,c_k|pa_j}) \\
&= \prod_{pa_j}\prod_{c_k} \frac{\lambda_{m,c_k|pa_j}^{N_i(c_k|pa_j)} e^{-\lambda_{m,c_k|pa_j}}}{N_i(c_k|pa_j)!} \qquad (4.4)
\end{aligned}$$

EM algorithm: We use Expectation-Maximization (EM) algorithm to infer the parameters [18]. In the Expectation-step, we use the current parameter estimate $\Theta^{(i-1)}$ to find the posterior probability $p(\mathbf{Y}|\mathbf{X}, \Theta^{(i-1)})$.

$$p(\mathbf{Y}|\mathbf{X}, \Theta^{(i-1)}) = q_{i,m} \propto \alpha_m . \prod_{pa_j}\prod_{c_k} p(N_i(c_k|pa_j)|\lambda_{m,c_k|pa_j}) \text{ subject to } \sum_{m=1}^{M} q_{i,m} = 1$$

Next, we use this posterior distribution to find the expectation of the complete-data log likelihood $Q(\Theta, \Theta^{(i-1)})$. In general, we have

$$Q(\Theta, \Theta^{(i-1)}) = \sum_{\mathbf{Y}} p(\mathbf{Y}|\mathbf{X}, \Theta^{(i-1)}) . \log p(\mathbf{X}, \mathbf{Y}|\Theta) = \sum_{m=1}^{M}\sum_{i=1}^{N} q_{i,m} . \log(\alpha_m . p(\mathbf{x_i}|\theta_m))$$

Here, $\mathbf{X}$ and $\Theta^{(i-1)}$ are constants, Θ is a variable that we wish to adjust and $\mathbf{Y}$ is a random variable governed by the distribution $p(\mathbf{Y}|\mathbf{X}, \Theta^{(i-1)})$. For the Bayesian mixture of Poissons, we obtain the complete-data log likelihood as,

$$Q(\Theta, \Theta^{(i-1)}) = \sum_{m=1}^{M}\sum_{i=1}^{N} q_{i,m} \Bigg(\log(\alpha_m)$$

$$+ \quad \sum_{pa_j} \sum_{c_k} (N_i(c_k|pa_j) \log \lambda_{m,c_k|pa_j} - \lambda_{m,c_k|pa_j}) \Bigg) \tag{4.5}$$

subject to the constraint,$\sum_{m=1}^{M} \alpha_m = 1$. In the Mazimization-step, we determine the new parameter estimate Θ by maximizing this function.

$$\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)})$$

The maximum likelihood estimates for the Bayesian mixture of Poissons are,

$$\alpha_m = \frac{\sum_{i=1}^{N} q_{i,m}}{N} \ , \ \lambda_{m,c_k|pa_j} = \frac{\sum_{i=1}^{N} q_{i,m}.N_i(c_k|pa_j)}{\sum_{i=1}^{N} q_{i,m}} \tag{4.6}$$

To initialize the EM algorithm, we randomly assign each read to a cluster m. The posterior probability $q_{i,m}$ is set to 1, if read i is assigned to cluster m and 0 otherwise. We then proceed with the M-step. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function.

4.2.2 Two-Way Bayesian Mixture of Poissons

Higher order words are known to be more discriminative than shorter ones[63]. However, with the increase in the length of words, the length of the read vector grows exponentially (e.g, for $l = 10, 4^l \approx 10^6$). Moreover, many words will tend to similar distributions and hence, can be clustered together into a "word group". At the same time, the number of distinct words in any read is usually substantially smaller than the number of dimensions. That is, the feature matrix becomes high-dimensional and sparse. Hence, the model may fail to predict the true distribution of different components. Therefore, dimension reduction becomes necessary before estimating the components in the model. However, reduction of the number of words using feature selection cannot be too aggressive, otherwise the clustering accuracy will suffer.

We handle the above challenge by "word grouping" that was described in the previous chapter. The cluster means are regularized by dividing the words into groups and then constraining the parameters for the words within the same

group to be identical. The grouping of the words is not pre-determined, but optimized as part of the model estimation. This implies that for every group, only one statistic for all the words in this group is needed to cluster reads. Note that we make a distinction on the use of "cluster" to refer to binning of reads sampled from the same species and "group" to refer to binning of words within a cluster. The parameter constraint implies that words have the same distribution within each cluster, while for reads in different clusters, words may follow different distributions. For simplicity, we assume that all clusters have the same number of word groups.

In the Bayesian mixture of Poissons, with the increase in l, the number of Poisson parameters to be estimated increases exponentially. We group the set of Poisson parameters corresponding to each parent into its Poisson vector. Therefore, we have p different Poisson vectors corresponding to p configurations of parents. We divide the parents into groups and constrain the Poisson vector distributions corresponding to parents within the same group to have identical parameters. The grouping of the parents is optimized as part of model estimation. Let L be the number of groups within each cluster. Let $c(m, pa_j) \in 1, 2, ..., L$ denote the group that parent pa_j belongs to in class m. All parents belonging to a group l, have Poisson parameter $\lambda_{m,c_k|l}$. Let the number of parents in group l of class m be η_{ml}.

$$p(\mathbf{x_i}|\lambda_{\mathbf{m}}) = \prod_{pa_j} \prod_{c_k} \frac{(\lambda_{m,c_k|l}^{N_i(c_k|pa_j)} e^{-\lambda_{m,c_k|l}})}{N_i(c_k|pa_j)!} \text{ where } c(m, pa_j) = l \tag{4.7}$$

Now, we can perform clustering using no more than order of ML dimensions. Word grouping leads to dimension reduction in this precise sense. We can derive an EM algorithm similar to the one outlined above to estimate the Multinomial parameters $\theta_{m,c_k|l}$ where $m \in 1, ..., M, l \in 1, ..., L$, the group assignment function $c(m, pa_j) \in 1, ..., L$, where $m \in 1, ..., M, j \in 1, ..., p$ and the proportion of mixture components $\alpha_m, m \in 1, ..., M$. We update the parameters as given below:

$$\alpha_m = \frac{\sum_{i=1}^{N} q_{i,m}}{N} \text{ , } \lambda_{m,c_k|pa_j} = \frac{\sum_{i=1}^{N} q_{i,m} . \sum_{pa_j \in l} N_i(c_k|pa_j)}{\eta_{ml} \sum_{i=1}^{N} q_{i,m}} \tag{4.8}$$

Once $\theta^{(t+1)}_{m,c_k|pa_j}$ is fixed, the word cluster index $c^{(t+1))}(m,j)$ can be found by doing a linear search over all components:

$$c(m, pa_j) = \arg\max_l \sum_{i=1} q_{i,m} \sum_{c_k} (x_{ij} \log(\lambda_{m,c_k|l}) - \lambda_{m,c_k|l})) \tag{4.9}$$

4.2.3 Bayesian Mixture of Multinomials

Theorem: If $X = (X_1, X_2, .., X_p)$ are independent Poisson variables with parameters, $\lambda_1, \lambda_2, .., \lambda_p$ respectively, then the conditional distribution of $(X_1, X_2, .., X_p)$ given that $X_1 + X_2 + ... + X_p = n$ is multinomial with parameters λ_j/λ, where $\lambda = \sum \lambda_j$, i.e. $Mult(n, \pi)$, where $\pi = (\lambda_1/\lambda, \lambda_2/\lambda, ..., \lambda_p/\lambda)$.

The above theorem implies that the likelihood based inferences about π are the same whether we regard $X_1, X_2, .., X_p$ as sampled from p independent Poissons or from a single multinomial. In our case, n refers to the length of the reads and our interest lies in the proportion of words in the reads. Any estimates, tests. inferences about the proportions will be the same whether we regard n as random or fixed. In the Poisson model, we regard the different word counts to be independently distributed, whereas, in the multinomial model the words counts are regarded as being multinomially distributed and hence are correlated.

We now derive the Bayesian mixture of Multinomials as standardized Bayesian mixture of Poissons. We assume that the distribution of words within the reads of a species is governed by the parameters of a multinomial distribution $\mathbf{\Theta}$. Let $P_m(c_k|pa_j) = \theta_{m,c_k|pa_j}$ be the probability of parent word pa_j followed by c_k, in reads belonging to species m. We have $\sum_{c_k \in (A,C,T,G)} N_i(c_k|pa_j) = N_i(pa_j) \ \ \forall pa_j$. The sum of CPDs is well-defined, $\sum_{c_k \in (A,C,T,G)} \theta_{m,c_k|pa_j} = 1 \ \forall m$ and $\forall pa_j$ Each local species distribution is a collection of multinomial distributions, one for each configuration of pa_j. $\theta_{\mathbf{m}} = \{((\theta_{m,c_k|pa_j})_{\forall c_k})_{\forall pa_j}\}$. Therefore, within every species m, for each configuration pa_j of parents, we get an independent multinomial problem, $Mult(\theta_{\mathbf{m,c|pa_j}}) = (\theta_{m,c_k|pa_j})_{\forall c_k}$, that has the same parameters across reads of a species.

$$\mathbf{\Theta} = \{\theta_{\mathbf{m}} : \forall m \in 1, .., M\}$$
$$\theta_{\mathbf{m}} = \{\theta_{\mathbf{m,c|pa_j}} : \forall pa_j \in \{pa_j^1, pa_j^2, .., pa_j^p\}\}$$

$$Mult(\theta_{\mathbf{m,c|pa_j}}) = \theta_{\mathbf{m,c|pa_j}} = \{\theta_{m,c_k|pa_j} : \forall c_k \in (A,C,T,G)\} \quad (4.10)$$

Therefore, the likelihood of the data will be,

$$p(\mathbf{x_i}|y_i = m) = p(\mathbf{x_i}|\theta_{\mathbf{m}}) = \prod_{pa_j} \prod_{c_k \in (A,C,T,G)} \theta_{m,c_k|pa_j}^{N_i(c_k|pa_j)} \quad (4.11)$$

EM algorithm: To initialize, we randomly assign each read to a cluster m. The posterior probability $q_{i,m}$ is set to 1, if read i is assigned to cluster m and 0 otherwise. We proceed with the M-step.

1. **E-Step:** We estimate the posterior probability $q_{i,m}$ as,

$$q_{i,m} \propto \alpha_m . \prod_{pa_j} \prod_{c_k \in (A,C,T,G)} \theta_{m,c_k|pa_j}^{N_i(c_k|pa_j)} \text{ subject to } \sum_{m=1}^{M} q_{i,m} = 1 \quad (4.12)$$

2. **M-Step:** The M-Step uses $q_{i,m}$ to compute the expectation of complete data log likelihood, taking into account the constraints.

$$Q(\Theta, \Theta^{(i-1)}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \left(q_{i,m} . \log(\alpha_m) + q_{i,m} . \sum_{pa_j} \sum_{c_k} N_i(c_k|pa_j) \log \theta_{m,c_k|pa_j} \right) + \beta(\sum_{m=1}^{M} \alpha_m - 1) + \sum_{m} \sum_{pa_j} \gamma_{m,pa_j} (\sum_{c_k} \theta_{m,c_k|pa_j} - 1)$$

The maximum likelihood estimates of parameters,

$$\begin{aligned} \alpha_m &= \frac{\sum_{i=1}^{N} q_{i,m}}{N} \\ \theta_{m,c_k|pa_j} &= \frac{\sum_{i=1}^{N} q_{i,m} . N_i(c_k|pa_j)}{\sum_{i=1}^{N} \sum_{c_k} q_{i,m} . N_i(c_k|pa_j)} = \frac{\sum_{i=1}^{N} q_{i,m} . N_i(c_k|pa_j)}{\sum_{i=1}^{N} q_{i,m} . N_i(pa_j)} \end{aligned} \quad (4.13)$$

The multinomial model conditions on the total count of parents, therefore, some clustering information is lost.

4.2.4 Two-Way Bayesian Mixture of Multinomials

For each species m, we have p different multinomial distributions, corresponding to p configurations of parents. We divide the parents into groups and constrain the multinomial distributions corresponding to parents within the same group to have identical parameters. The grouping of the parents is optimized as part of the model estimation. Let L be the number of groups within each cluster. Let us define a function $c(m, pa_j) \in 1, 2, ..., L$, which tells us the group that parent pa_j belongs to in class m. All parents belonging to a group l, have multinomial parameter $\theta_{m,l} = \theta_{m,c(m,pa_j)}$. The group assignments of the parents vary from cluster to cluster. Let the number of words in group l of class m be η_{ml}.

$$P_m(c_k|pa_j) = \theta_{m,c_k|pa_j} = \theta_{m,c_k|l} \text{ where } c(m, pa_j) = l \tag{4.14}$$

The likelihood of $\mathbf{x_i}$ now becomes:

$$p(\mathbf{x_i}|\theta_\mathbf{m}) = \prod_{pa_j} \prod_{c_k} \theta_{m,c_k|l}^{N_i(c_k|pa_j)} \text{ where } c(m, pa_j) = l \tag{4.15}$$

We can derive an EM algorithm similar to the one outlined above to estimate the Multinomial parameters as,

$$\alpha_m = \frac{\sum_{i=1}^{N} q_{i,m}}{N} \tag{4.16}$$

$$\theta_{m,c_k|pa_j} = \frac{\sum_{i=1}^{N} q_{i,m} \cdot \sum_{pa_j \in l} N_i(c_k|pa_j)}{\sum_{i=1}^{N} \sum_{pa_j \in l} \sum_{c_k} q_{i,m} . N_i(c_k|pa_j)} = \frac{\sum_{i=1}^{N} q_{i,m} \cdot \sum_{pa_j \in l} N_i(c_k|pa_j)}{\sum_{i=1}^{N} q_{i,m} \cdot \sum_{pa_j \in l} N_i(pa_j)}$$

Once $\theta_{m,c_k|pa_j}^{(t+1)}$ is fixed, the word cluster index $c^{(t+1))}(m, j)$ can be found by doing a linear search over all components:

$$c(m, j) = \arg\max_l \sum_{i=1} \sum_{c_k} q_{i,m} N_i(c_k|pa_j) \log \theta_{m,c_k|l} \tag{4.17}$$

4.3 Results

4.3.1 Datasets

Metagenomics being a relatively new field, lacks standard datasets for the purpose of testing clustering algorithms[41]. As the "true solution" for sequence data generated from most metagenomic studies is still unknown, we focus on synthetic datasets for benchmarking. We use Metasim to simulate metagenomes[27]. It takes as input the sequencing technology to be used, a set of known genomes, length of the reads and an abundance profile that determines the relative abundance of each genome in the simulated dataset. The genomes used for generating the synthetic metagenomes were downloaded from National Center of Biotechnology Information (NCBI). We generated datasets with read lengths between 50 and 1000 bps and various abundance ratios.

The algorithms were implemented in Matlab. The space and time complexity scale linearly with the number of reads and species and quadratically with the number of dimensions in the search space. Our methods converged for all the cases we tested and was robust to the choice of initial conditions. In practice, a 2-species dataset containing 10,000 reads with word length of 4 and read length of 500 bps took approximately 2 minutes to run on an Intel Core 4- Duo processor.

In order to assess the robustness of our proposed method, we tested it across datasets representative of δ^* values (defined in the last chapter) ranging from 34 to 340. In general, lower δ^* values correspond to "closely related species" and higher values correspond to "distant species". In order to calculate the clustering accuracy, we assign each cluster to the source species that is most frequent in the cluster. Accuracy is given by the percentage of correct read assignments.

The number of species in each dataset is supplied as an input. Determining the number of clusters from a statistical perspective is a difficult problem[65]. Maximum likelihood favors more complex models leading to over-fitting and hence is unable to address this issue. Previously, 16s/18s rDNA have been used for phylotyping and assessing species diversity using a rare-fraction curve. Most methods rely on heuristics to guide the choices of clusters. Determining species diversity is still an active area of research and we do not address it in this chapter.

In Figure 4.2, we compare the performance of all four probability models

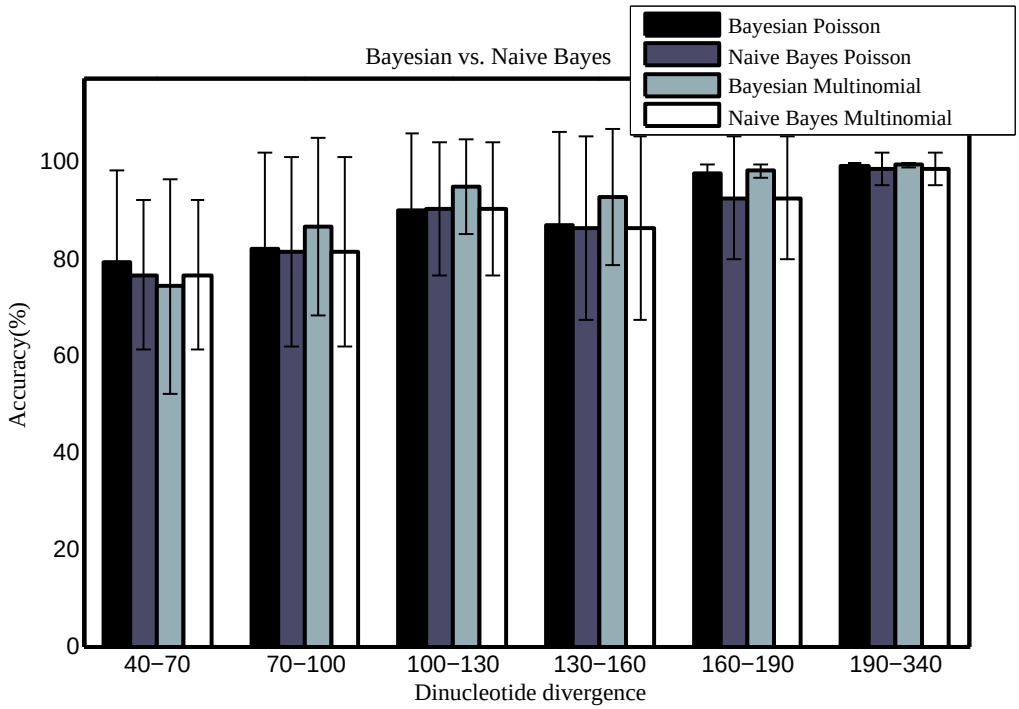


Figure 4.2. Comparison of performance of Bayesian mixture of Poissons and Multinomials with their Naive Bayes counterpart for 2-species dataset with δ^* (measured on 50-kb contigs) values ranging from 34 to 340, with 34 corresponding to "closely similar species" and 340 to "very distant species". We used a word length of 4.

proposed in the chapter. We measured accuracy over 400 datasets for δ^* values ranging from 34 to 340. We used a word length of 4. We observe a positive correlation between δ^* and the accuracy of our methods, as also noted in [33]. Bayesian networks efficiently encode the conditional dependencies between the words due to overlaps. Both the Bayesian mixture of Multinomials and Poissons have a better clustering accuracy than their Naive Bayes counterparts. Though, Naive Bayes methods offer a much simpler and faster alternative at the cost of a slightly lower performance. A possible justification for higher accuracy of multinomial models is that it regards the word counts as being multinomially distributed and hence captures the correlation between words counts.

We use the Bayesian Poisson mixture model for the remaining analysis. The results for Bayesian mixture of Multinomials were not substantially different unless otherwise stated. The efficacy of clustering methods depends on the number of species in the dataset, the read length and relative abundances of source

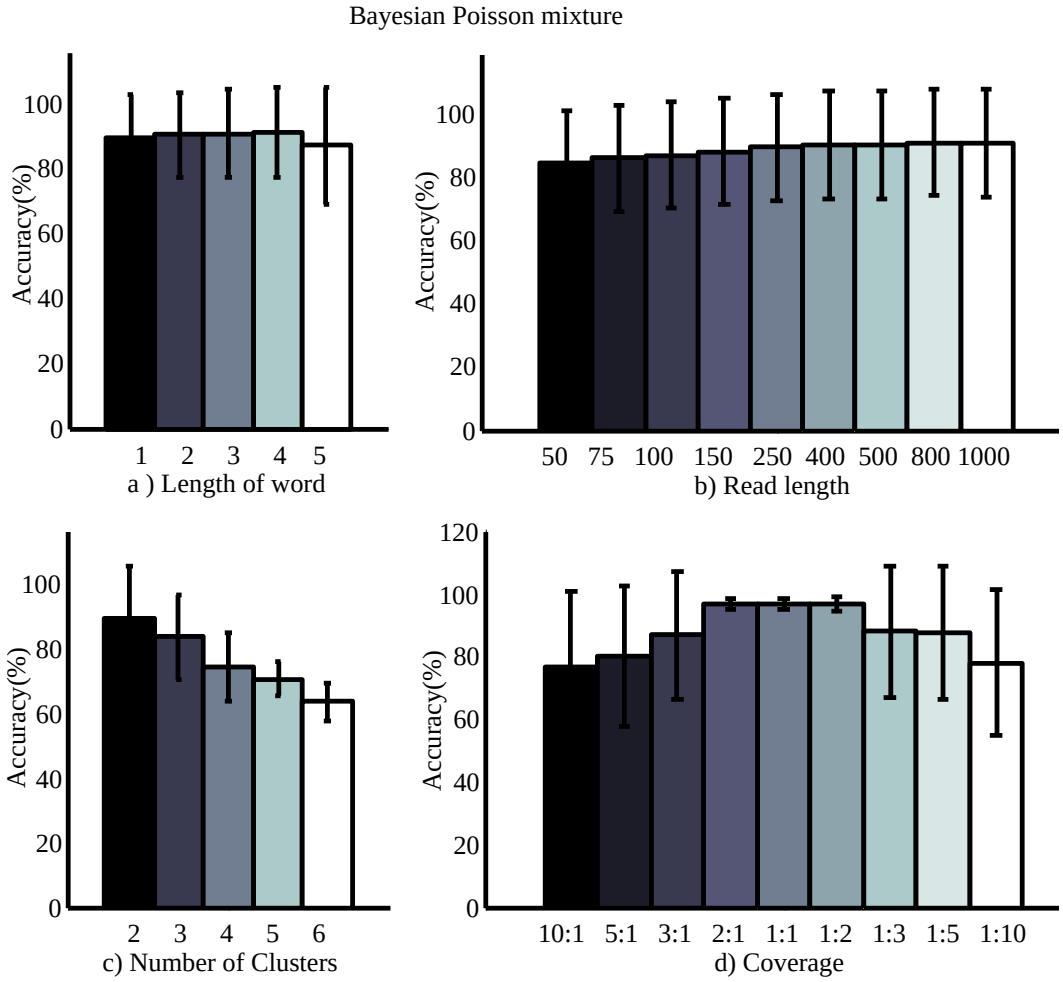


Figure 4.3. Performance of Bayesian mixture of Poissons a) Accuracy Vs. length of word b) Accuracy Vs. read length c) Accuracy Vs. number of clusters d) Accuracy Vs. coverage

genomes in the microbial community.

4.3.2 Accuracy Vs. Coverage

We systematically evaluated the robustness of our method to changes in the coverage ratio between species representative of various intergenomic differences. Binning results for 20 sets of simulated metagenomes with two species each is summarized in Figure 4.3. The datasets contained 10000 reads of length 500 bps each. We used a word length of 2. We varied the coverage ratio from 10:1 to 1:10 in stages, for the two species. From Figure 4.3, we note that there was only a

slight drop in performance for extreme coverages, when the fractional content of the species reduces to less than 10%. Therefore, the proposed method is suited for binning relatively rare species as well.

4.3.3 Accuracy Vs. Number of clusters

Next, we analyzed the accuracy and applicability of Bayesian mixture of Poissons in binning reads from low complexity communities, containing 2-6 species (see Figure 4.3). The results were averaged over 50 datasets of varying divergences. Given that the multi-species dataset may contain reads from species with little intergenomic differences, there was a slight degradation in performance with the increase in number of species. This is in agreement with the results from the 2 species dataset, considering that the total coverage of each species is much lower in a multi-species dataset. The results indicate that our method is suitable for binning reads belonging to dominant species, and that binning relatively rare species in a multi-species dataset may require modifications to the present Bayesian formulation.

4.3.4 Accuracy Vs. Read length

Binning results on 2-species dataset were evaluated using read length ranging from 50 bps to 1000 bps. The results are shown in Figure 4.3. We can infer that the classification accuracy is mostly correlated to the read length. A point worth noting is that the drop our method's error rate is bounded by 5% for datasets with read length as short as 50 bps when compared to that with read length of 1000 bps.

4.3.5 Accuracy Vs. Length of word

In general, the discriminative power of the models increases with the length of words, despite the increasing space complexity. In Figure 4.3, we plot the clustering accuracy of the Bayesian mixture of Poissons with increasing word length. The values have been averaged over 50 datasets of varying δ^* values. Notice that initially the accuracy increases with the word length. For word

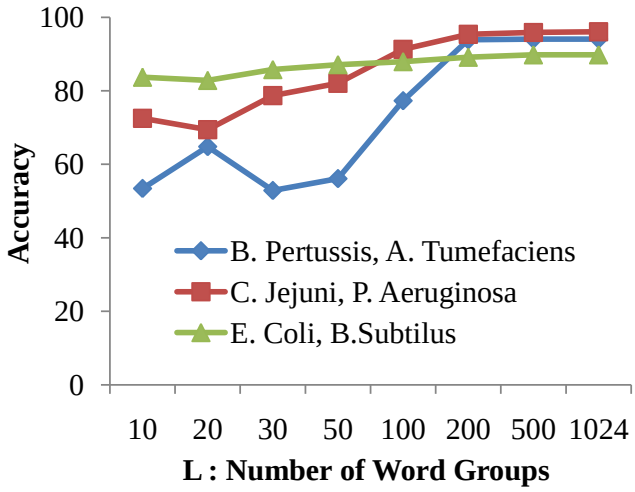


Figure 4.4. Performance of Two-way Bayesian Poisson mixture model for values of word groups, L, varying from 10 to 1024. A word length of 5 is used. Each dataset contained 50,000 reads of 500 bps each

lengths beyond five, the accuracy begins to drop. This is because the feature matrix becomes high-dimensional and sparse. Hence, the model fails to predict the true feature distribution of different components. This necessitates dimension reduction before estimating the components in the model.

In this chapter, we perform "word grouping" to handle the above challenge. We propose a two-way mixture model where the mixture clusters induce a partition of the reads as well as of words. We used word length of 5 and varied the number of word groups from 10 to 1024 in stages (Figure 4.4). Performance stabilizes close to its optimal value at $L = 100$. This implies that the data can be classified using no more than ML dimensions, a significant reduction from the original number of dimensions. That is, the characteristic vectors are of a much lower dimension. Note that it is difficult to know a priori, the exact value of L that yields the best clustering. However, among the values we tested, lower values of L provided a higher accuracy.

Finally, in Figure 4.5, we compare the accuracy of our proposed Gaussian mixture model with state-of-art unsupervised composition-based method Scimm on several datasets[32]. We varied the δ^* values from 60 to 300. We used a read length of 200 bps and word length of 4. We averaged results over 60 randomly chosen datasets. As the number of dimensions is relatively small, our method

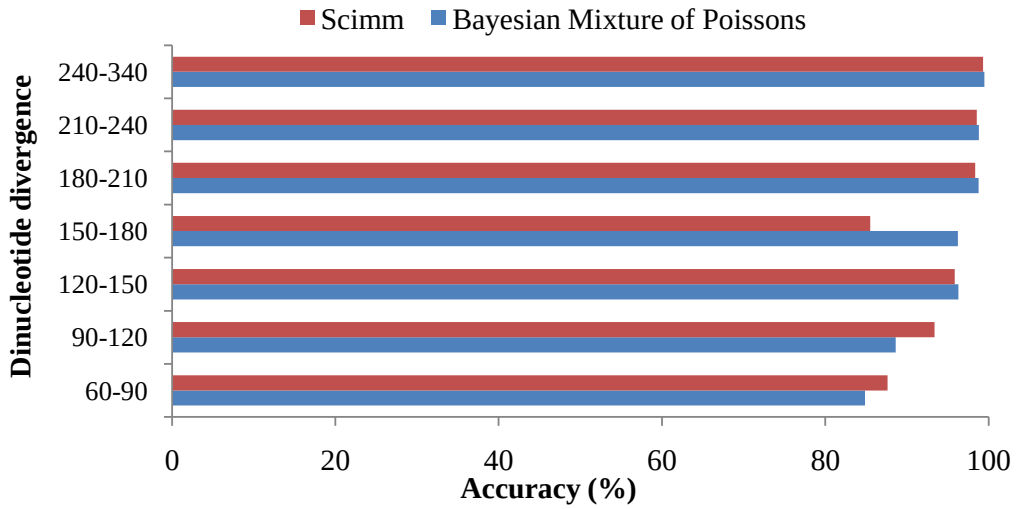


Figure 4.5. Comparison of performance of Bayesian mixture of Poissons with Scimm. We varied the δ^* values from 60 to 300. We used a read length of 200 bps and word length of 4.

performs well without word grouping too. For δ^* values of 150 and above, our method performs marginally better than Scimm. Though for δ^* values below 90, Scimm does better than our method.

4.4 Conclusion

In this chapter, we propose multivariate Bayesian methods based on Poisson and Multinomial mixture model to cluster the reads in a metagenome by their species of origin. This work demonstrates the use of statistically based mixture models for analysis of metagenome datasets by suitable choices of probability distributions. The Poisson and Multinomial models can effectively cluster the reads when the word counts are very low. An additional reason to use these distributions is their flexibility, stability and ease of parameter estimation. Bayesian networks are used to represent the conditional dependencies between the words. We examine the sensitivity of the method to the number of species, abundance profile and length of reads within the dataset. Much work needs to be done to validate the usefulness of these model for real metagenome datasets. Our method is an unsupervised method that does not require any training data. This is critical for success as most metagenomic datasets contain reads from unexplored phyla

which cannot be labeled into one of the existing classes. However, we still need to specify the number of species for the algorithm. A future direction for our work is to overcome this limitation. Our framework complements the existing similarity-based and abundance-based methods and hence, can be combined with such methods to obtain a better performance. We intend to develop such hybrid methods in the future that can tackle the problem of classifying sequences in complex metagenomic communities. Our probabilistic approach can be used to identify reads which belong to more than one species and occlude the cluster boundaries. Such reads should be further investigated to identify the presence of conserved regions. The methods have been tested on metagenome sequence data, but can be adapted for use with a variety of discrete sequence data. We have not analyzed the methods on other discrete sequence data such as document clustering data, web-logs, purchase history or stock market data among others.

Chapter 5

Mutant Bin for Viral Haplotype Estimation

5.1 Background and Motivation

Next Generation Sequencing (NGS) technologies generate data more efficiently, economically and with a greater depth than ever before. NGS has opened up an array of possibilities for many applications including whole-genome sequencing, epigenetics, metagenomics and characterization of Pathogen. Of these, the characterization of genetic diversity in heterogeneous pathogens such as viral populations has recently gained significant interest. Although a host of methods for whole genome assembly have been developed, reconstruction of heterogeneous populations using NGS data still remains a challenge. As compared to existing technologies, reads produced by NGS are typically shorter and more error-prone. Thus, many computational challenges arise while analyzing deep sequence data from heterogeneous populations[48].The computational method we present here aims to quantify the genetic diversity within a heterogeneous population based on a set of deep sequencing reads.

At any given time, within-host virus populations consist of a collection of distinct, albeit closely related genetic variants that are known as Quasispecies. Each individual variant, also known as haplotype, occurs with a different relative frequency, with some over a hundred times less abundant than the dominant

variants. The high genetic diversity of a pathogen population has important consequences in disease progression as it allows the virus to evolve rapidly, and evade host defenses and therapeutic interventions. The high coverage and enormous sequence data output by NGS technologies has the potential to resolve the genetic variation within the sample and thereby infer the population structure[25]. This application is not only a promising approach for study of disease progression, detection of emerging drug resistant variants and vaccine design, but also would be useful for studying evolution of viral populations.

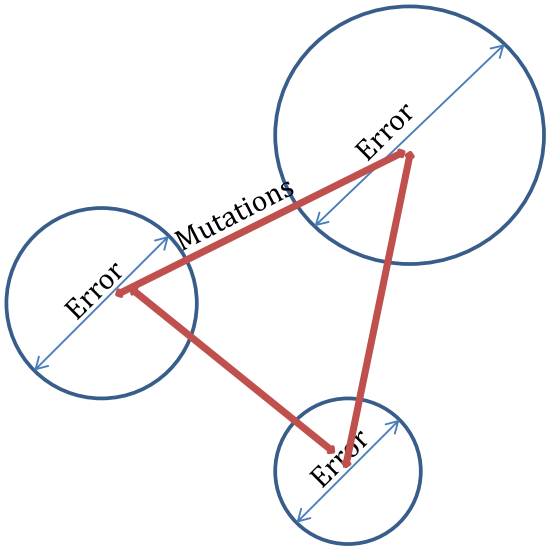


Figure 5.1. Virus Haplotype model in 2-D sequence space. Variation within cluster is due to errors in sequencing. Variation between clusters is due to haplotype differences.

A number of methods have been published for quasispecies reconstruction that are able to infer genomes of individual haplotypes as well as their prevalence. In [20] and [73], the authors proposed a set of methodologies based on a graph theoretic solution, applying clustering and Bayesian based error correction as preprocessing. A set of single haplotypes of the quasispecies were obtained by constructing an overlap graph of non-redundant, aligned reads and by calculating a minimal coverage set of paths over the graph. In [49], a reconstruction algorithm based on combinations of multinomial distributions was designed to take into consideration the overlaps between reads with similar frequencies. A few other combinatorial algorithms have been proposed that extract a minimal subset of

haplotypes that explain all observed reads[3, 29]. These methods were mostly applied to high diversity (3-10%) HIV datasets. Many recent studies, however, have sequenced data from populations of eukaryotic individuals that represent less genetic diversity (0.6-2%)[45]. The above methodologies require the presence of a reference genome, which drastically limits the number of species that can benefit from these new technologies. *De novo* assembly of short reads is, as such, an NP-hard problem. Moreover, the presence of rampant structural polymorphisms, high mutation rate and sequencing artefacts makes it difficult to obtain a consensus sequence. In [5], the authors explored the limitations of consensus sequencing. In some regions of HIV *env* gene (eg. V1, V2, V4, V5), where insertions and deletions accumulate rapidly, the consensus sequence has no biological meaning, making their use for phylogenetic analyses questionable. To the best of our knowledge, not much has been published about resolving the population structure when an assembled reference genome is lacking or is not well-defined.

We present a computational framework, Mutant-Bin, for clustering individual haplotypes in a viral population and determining their prevalence, based on a set of deep sequencing reads. The method when applied to simulated data sequenced at a high coverage, clustered a high percentage of true haplotypes with low false positive rates, even at a low genetic diversity and estimated relative frequencies that are in agreement with the true proportions. The main advantages of our method are that: (i) it enables determination of the population structure and haplotype frequencies when a reference genome is lacking, and hence, avoids the costly clonal sequencing and the compute intensive alignment required by other methods; (ii) the number of haplotypes does not have to be specified in advance; (iii) it identifies the regions with co-occurrence of polymorphic sites in a subset of haplotypes and the frequency with which they appear in the population. Phylogenetic and evolutionary studies that do not require the knowledge of the sequence itself, but only the number of variable sites and the nucleotides in these sites can benefit from such a method. Our main motivation is to use this framework to survey the genetic diversity of viral populations in situations, where a reference genome is not well defined.

We test Mutant-Bin on simulated and real deep sequencing datasets representing diversities between $0.1 - 10.0\%$. We investigate the performance of our

method as a function of pairwise distances between haplotypes, sequencing errors, relative frequencies and number of haplotypes. We assess its performance in isolating regions containing polymorphic sites into clusters and estimation of haplotype frequencies. Also, we compare our method to the existing state-of-the-art, ShoRAH[73].

5.2 Methods

Consider a quasispecies with K haplotypes of genome length G that appear with a frequency of $\mathbf{x} = \{x_k\}_{k=1}^{K}$ in the population. The number of haplotypes K and their frequencies are unknown. Let $\mathbf{R}$ be the set of reads obtained from the viral population. The fraction of reads sequenced from haplotype k will be proportional to its frequency x_k. Our objective is to identify regions within each haplotype that contain the polymorphic sites, cluster individual haplotypes and to infer their frequency in the population.

We define the pairwise difference d_{ij} between two sequences s_i and s_j as the hamming distance between them, i.e., the number of base positions on which the two sequences differ. The diversity of a population containing K haplotypes is then,

$$D_{pop} = \frac{\sum_{i=1}^{K-1}(\sum_{j=i+1}^{K} d_{ij})}{K(K-1)/2}$$

We designate as the reference, the variant with the lowest average pairwise difference in the population. The regions within the reference genome that have polymorphic sites with derived nucleotides (as opposed to ancestral) that result in different haplotypes are called suspect regions (see Figure 6.1).

Mutant-Bin is an application of the Lander-Waterman model to viral population estimation and is based on the l-tuple content(ordered sequence of length l) of the reads . The framework for Mutant-Bin consists of three steps. First, the distribution of l-tuples within the population is modeled as a mixture of Poisson distributions. The means of Poisson distributions correspond to the coverage of suspect regions, determined by the subset of haplotypes in which the polymorphic sites co-occur. Second, the l-tuples are clustered by their frequency using the Variable-Bandwidth Mean-Shift algorithm (VBMS). We bin the l-tuples by their

cluster centers to uniquely identify the l-tuples spanning the polymorphic sites that co-occur in a subset of haplotypes and determine the frequency with which they appear. Finally, we propose a greedy heuristic to map the l-tuples to the genomes they originate from and thereafter infer the local haplotype structure.

5.2.1 Mixture of Poisson Distributions

According to Lander-Waterman model, the probability that a base is sequenced m times follows a Poisson distribution, $P(m) = \lambda^m \frac{e^{-\lambda}}{m!}$, where λ is the coverage (number of bases sequenced per position) of the experiment[34]. We define l such that all l-tuples appear at most once within the genome. Then, the number of occurrences of l-tuples in a set of reads also follows a Poisson distribution with parameter $N(L-l+1)/(G-L+1) \approx NL/G$, where N is the number of reads, L is the length of reads and G is the genome length.

In a viral population, at any given time each haplotype occurs with a different relative frequency. In Figure 5.2, the frequency spectrum illustrates how the distribution of l-tuples within a viral quasispecies can be modeled by a mixture of Poissons for a population containing two haplotypes that appear with a coverage of x_A and x_B (which is proportional to their relative frequencies in the population)[1]. The figure depicts two genomes of length G and a shaded region that corresponds to the base positions on which the two genomes differ (i.e., the suspect region with polymorphic sites). The distribution of l-tuples that span the portion of the genomes that is common to both haplotypes is approximated by a Poisson distribution with mean $x_A + x_B$ and those that span the suspect regions unique to A and B with means x_A and x_B respectively. Now the problem of identifying different suspect regions is transformed to that of modeling mixture of Poisson distributions.

The underlying basis of the binning procedure is that the l-tuples sampled from suspect regions will appear with a low count relative to those sampled from the rest of the genome. By determining the number of times each l-tuple occurs and the fraction of the genome it accounts for, we can uniquely identify the tuples belonging to the suspect regions (Figure 5.4). Typically, if we have a high coverage

[1]The frequencies of the haplotypes in the population can be approximated by their coverages. Hence, we use the terms *frequency* and *coverage* interchangeably.

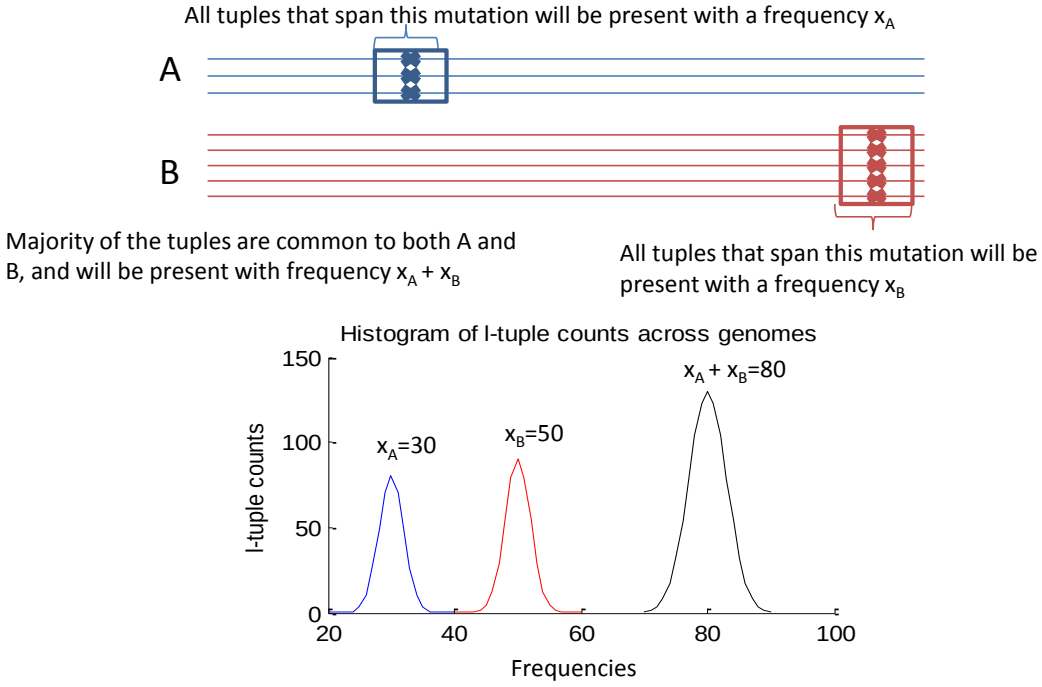


Figure 5.2. Virus population with two haplotypes in frequencies, x_A and x_B. Frequency spectrum of l-tuple counts in the population. The figure depicts two genomes of length G and a shaded region that corresponds to the base positions on which the two genomes differ (i.e., the suspect region).

and a good separation between the Poisson means, we can distinguish the suspect regions, no matter how similar they are. Deep sequencing technologies make it possible to sample viral populations at a great depth.

The frequencies $\{x_i\}_{i=1}^{K}$ with which the K haplotypes appear in the viral population are called *basic frequencies*. A polymorphic site at a given position can occur in one or more of K haplotypes. For instance, polymorphic sites for which exactly one haplotype in the population has a derived nucleotide will be most common, followed by those for which exactly two haplotypes in the sample have a derived nucleotide and so on[28]. The distribution of l-tuples that span the suspect region with polymorphic sites that co-occur in r haplotypes, say $(x_{j_1}, x_{j_2}, .., x_{j_r})$ will be Poisson with mean $y_i = \sum_{k=1}^{r} x_{j_k}$. We denote the set of such l-tuples by T_{y_i}. The maximum number of Poisson distributions that can be obtained is $2^K - 1$, though one usually does not observe all of them in a given population. Let us assume that we obtain M different Poisson distributions

with means $\mathbf{y} = \{y_i\}_{i=1}^{M}$ in the given population. We call these the *composite frequencies*. Our immediate goal is to determine the mean values of the different Poisson distributions that correspond to the *composite frequencies* and cluster the l-tuples using it. We note that a similar method was used by [38] for estimating the repeat content within a genome. The proposed method was recently adapted by [72] to classify reads within a metagenomic sample.

5.2.2 Cluster l-tuples using Variable Bandwidth Mean Shift Analysis

The count of l-tuples sampled from the viral population forms a mixture of Poissons. The frequency spectrum is multi-modal with modes corresponding to the frequency of suspect regions unique to each subset of haplotypes. We propose to use Variable Bandwidth Mean-Shift (VBMS) method to assign each l-tuple, based on its count, to a cluster which represents a mode of the Poisson distribution, using the adaptive *kernel density estimate* (KDE). The mean shift procedure was initially described in [15, 16] and adapted in [75] to detect and remove sequencing errors prior to assembly. Mean shift analysis is a robust non-parametric estimator of density gradient that estimates the modes of the distribution instead of the means. This technique is attractive, since it needs no prior knowledge of the number of clusters or the shape of the distribution. Moreover, a mode detection algorithm will not be affected by the variance in data, introduced due to presence of sequencing biases.

Given the set of reads $\mathbf{R}$ sequenced from the viral population, the algorithm starts by counting l-tuples in it (see Algorithm 1). Let $\{c_i\}_{i=1}^{L}$ denote the count of different l-tuples in $\mathbf{R}$, L being the total number of possible l-tuples. The *multivariate kernel density estimate* with kernel $K(\mathbf{x})$ for l-tuple k is defined as:

$$\tilde{f}(c_k) = \frac{a}{L} \sum_{i=1}^{L} \frac{1}{h_i} K(\| \frac{c_k - c_i}{h_i} \|^2) \tag{5.1}$$

where a is the normalization constant and h_i is termed the *kernel bandwidth*, and determines the range of influence of the kernel located at l-tuple i. Here, $\{c_i\}_{i=1}^{L}$ represent a random sample from some unknown density f. The kernel, $K(x)$,

is taken to be a spherically symmetric, non-negative function centered at zero and integrating to one. The adaptive bandwidth procedure estimates the density at each point c_i by taking the average of differently scaled kernels centered at each of the data points. For multivariate kernels, the optimum kernel yielding minimum mean integrated square error is the 1-d Epanechnikov kernel, with its profile defined as,

$$K_E(x) = \begin{cases} 1 - x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

While using the adaptive KDE $\tilde{f}(c_k)$ for clustering, our objective is to assign each data point to a cluster, based on the mode that point evolves to under a gradient ascent algorithm. The gradient of the kernel density estimate is:

$$\nabla \tilde{f}(x) = \frac{a}{L} \sum_{i=1}^{L} \frac{1}{h_i} \nabla K_E(\| \frac{x - c_i}{h_i} \|^2) \tag{5.3}$$

We define $g(x) = -K_E'(x)$. The mean shift vector is defined as an estimate of the normalized gradient of the underlying distribution and can be obtained from (4) as,

$$m(x) = \frac{\sum_{i=1}^{L} c_i g(\| \frac{x-c_i}{h_i} \|^2)}{\sum_{i=1}^{L} g(\| \frac{x-c_i}{h_i} \|^2)} - x$$

It has been proved that this process converges at a point where the estimate has zero gradient[15], i.e. the modes of the density. Therefore, an estimate of the normalized gradient can be obtained by computing the sample mean shift in a uniform kernel centered at c_k. Our method clusters l-tuples based on their corresponding mode at convergence. The obtained modes will correspond to the *composite frequencies*, $\mathbf{y} = \{y_i\}_{i=1}^{M}$. We represent the set of l-tuples clustered with mode y_i by C_{y_i}. Therefore, each cluster C_{y_i} contains the l-tuples that span the suspect region with polymorphic sites that co-occur in a subset of haplotypes such that $\sum_k x_k = y_i$.

Algorithm 1 Variable Bandwidth Mean Shift Analysis

Input: $\{c_i\}_{i=1}^{L}$, the count of different l-tuples in read set $\mathbf{R}$

Output: Modes $\mathbf{y} = \{y_i\}_{i=1}^{M}$ in the spectrum and $\{C_{y_i} = l\text{-tuples clustered with } y_i\}_{i=1}^{M}$

1. Compute the fixed bandwidth h_0 from the 1-dimensional plug-in rule proposed in [67].

2. Calculate the initial KDE $\hat{f}(c_x)$ of l-tuple k[61]:

$$\hat{f}(c_k) = \frac{a}{Lh_0} \sum_{i=1}^{L} K_E(\| \frac{c_k - c_i}{h_0} \|^2)$$

3. For each l-tuple c_i, compute its adaptive bandwidth, $h_i = h_0[\lambda/\hat{f}(c_i)]^{1/2}$, where $\log \lambda = n^{-1} \sum_{i=1}^{L} \log \hat{f}(c_i)$.

4. For each l-tuple k, initialize $c_k^{t=0}$ with c_k, the count of l-tuple k to be clustered.

   (a) Compute the mean shift vector $m(c_k^t)$ using equation (5)

   (b) Translate density estimation window: $c_k^{t+1} = c_k^t + m(c_k^t)$

   (c) Iterate above two steps until convergence.

5. l-tuples that converge to the same mode y_i form cluster C_{y_i}.

5.2.3 Cluster l-tuples using Expectation Maximization

In the case that we already know M, that is, the number of different composite frequencies in the dataset, we can use the Expectation-Maximization algorithm to determine the means of the Poisson distributions, $\mathbf{y} = \{y_i\}_{i=1}^{M}$ in the given population. For the sake of completeness, the method is outlined below.

Given the set of reads from a quasispecies, the algorithm begins by counting l-tuples in all the reads. Each composite frequency is mathematically represented by a parametric Poisson distribution with y_i equal to its mean coverage. Let the prior probability of number of reads sampled from composite frequency y_i be π_i.

The algorithm is as given below:

1. Initialize the prior probability π_i, and the composite frequencies y_i for $i = 1, 2..M$.

2. **Expectation-step:** Calculate the probability that the l-tuple $w_j, (j = 1, 2, ...L), L$ being the total number of possible l-tuples) coming from the i^{th} Poisson distribution, given its count c_j.

$$p(w_j \in i|c_j) = \frac{\pi_i}{\sum_{k=1}^{M} \pi_k (\frac{y_k}{y_i})^{c_j} e^{y_k - y_i}} \tag{5.4}$$

3. **Maximization-Step:** Calculate the new values of π_i and λ_i.

$$\begin{aligned} G &= \sum_{i=1}^{M} \sum_{j=1}^{L} p(w_j \in i|c_j) \\ \pi_i &= \frac{\sum_{j=1}^{L} p(w_j \in i|c_j)}{G} \\ y_i &= \frac{\sum_{j=1}^{L} c_j p(w_j \in i|c_j)}{G} \end{aligned} \tag{5.5}$$

4. Iterate steps 2 and 3 until the parameters converge or the number of runs exceeds the maximum number of runs.

Once the EM algorithm converges, we can estimate the probability of a read assigned to a bin, based on its l-tuple content.

5.2.4 Greedy Heuristic for Generating Set

Once we determine the *composite frequencies* $\mathbf{y} = \{y_i\}_{i=1}^{M}$, our next task is to find the *basic frequencies* $\mathbf{x} = \{x_i\}_{i=1}^{K}$, such that every y_i is the sum of a subset of $\mathbf{x}$ and the size of $\mathbf{x}$ is minimal. Such a set $\mathbf{x}$ is known as a *generating set* of $\mathbf{y}$. Determining the *generating set* is an NP-complete problem [14]. Using the greedy heuristic outlined below (Algorithm 2), the problem can be solved in polynomial time. Most of the algorithms for *generating set* in the literature are heuristic or approximative. Generating Set problem is related, among other

things, to planning radiation therapy. Work by Collins et al. provides some non-trivial lower bounds given certain constraints.

We propose a greedy heuristic for constructing the generating set $\mathbf{x}$ of $\mathbf{y}$. We do not allow $\mathbf{x}$ and $\mathbf{y}$ to be multisets. If some number x is repeated, then without loss of generality, we can replace x by $2x$ in the set. Let P_{y_i} be the representation of y_i, i.e. the subset of $\mathbf{x}$ such that $y_i = \sum_{x_k \in P_{y_i}} x_k$. Let D be the set of all possible differences in $\mathbf{y}$. The main idea is that while constructing $\mathbf{x}$, at each step, we choose the least y_i which does not already have a representation in $\mathbf{x}$. The condition $y_i \notin mode(D)$ ensures that we do not delete some x_r such that $x_p + x_q = x_r$ for some $x_p, x_q, x_r \in \mathbf{x}$.

In algorithm 2, $\mathbf{x}$ corresponds to the frequencies with which the haplotypes appear in the viral population, i.e. the *basic frequencies*. This implies, cluster C_{y_i} contains the l-tuples spanning the polymorphic sites that co-occur in haplotypes $x_k \in P_{y_i}$. We can now cluster the l-tuples by the *basic frequencies* of the genomes from which they originate. Initialize K bins, $\{B_k = \emptyset\}_{k=1}^{K}$. For each variant k, for all y_i, such that $x_k \in P_{y_i}$, we bin together the l-tuples from C_{y_i} into bin B_k. That is, for each *composite frequency* y_i and its representation P_{y_i}, define $\{B_k = B_k \cup C_{y_i}, \forall x_k \in P_{y_i}\}$. Ultimately, each B_k will consist of all l-tuples corresponding to variant k.

Definition 1 (Generating Set)**.** *Given a set of real numbers Y, is there a subset X of size K, such that Y is composed entirely of numbers that are sums of subsets of X.*

Definition 2 (3-SAT)**.** *Given a set of clauses $C_1, ..., C_k$, each of length 3, over a set of variables $X = (x_1, ..., x_n)$, does there exist a satisfying assignment.*

Theorem 3. *Generating Set is NP-complete.*

Proof. We first show that Generating Set problem is in NP. We assume that size of set $\mathbf{x}$ is polylogarithmic is size of $\mathbf{y}$. Given a set of numbers X, we can verify in linear time that all numbers in Y are indeed sum of subsets of X. We generate sums of one or more numbers from P and verify if we have covered every number in S.

We reduce 3-SAT to Generating Set problem[14]. We first assume that we are given a 3SAT formula with n variables and k clauses and that every clause in

our input formula has exactly three literals (or repeat literals in the same clause to make this true). In our notation, we represent the numbers in decimal, with a column for each of the n variables and a column for each clause in the formula. We introduce an item a_i for each of the $2n$ literals. This item will have a 1 in the column for its variable, a 1 in the column of each clause where the literal appears, and zeros elsewhere. We also have two items for each clause, each with a 1 in the column for that clause and zeros elsewhere. We add a target number that has a 1 for each variable column and a 3 for each clause column into the set.

We now have to prove that there is a basic subset of size $2n+2k$ iff the formula is satisfiable. If there is a satisfying assignment, we select the item for each literal in that assignment. This has one 1 in each variable column, and somewhere from one to three 1s in each clause column. Using extra items as necessary, we can reach each target in set S. Each of the remaining numbers are present in the subset as is.

Conversely, if we find a basic subset, then there exist a subset of numbers that add to the target. We must have chosen one item with a 1 in each variable column, so we have picked n variables forming an assignment. Since we have three 1s in each clause column and at most two came from the extra items, we must have at least one 1 in each clause column from our assignment, making it a satisfying assignment. □

5.2.5 Inferring Phylogeny

Given a viral population of K haplotypes, the frequency spectrum represents the distribution of polymorphic sites with derived nucleotides that co-occur in a subset of haplotypes. Given how mutations occur and how a population evolves, it is possible to predict the shape of the spectrum in theory[28]. Using our method, we can work backwards to infer the population parameters and the evolutionary tree, given the shape of the spectrum. Such analyses play an important part in glsPhylogeneticsphylogenetics. The graphical structure of the tree is intrinsic to the co-occurrence of polymorphic sites within haplotypes. It is possible to derive the tree and branch lengths, given such a spectrum. The polymorphic sites that have derived nucleotides, co-occurring in a subset of haplotypes $x_k \in P_{y_i}$ will

appear with *composite frequency* y_i in the frequency spectrum. The number of such polymorphic sites shared between haplotypes in subset P_{y_i} will be proportional to the relative area of the Poisson curve with frequency y_i in the spectrum. Thus, one can infer the mutation probabilities $\mathbf{p}$ from the spectrum and then use $\mathbf{p}$ to infer the tree. Therefore, it is possible to construct the evolutionary tree without having to deal with the complications of the intermediate assembly step. Figure 5.2.5 shows an illustration of the entire methodology of Mutant-Bin for an example dataset containing three haplotypes in the ratio $1:3:5$.

5.3 Experimental Results

We evaluated our method on simulated as well as real deep sequencing data from HIV samples using 454/Roche FLX technology. The first 2 kbp of HIV-1 genome was the starting point for all our simulations. We used Metasim's population sampler to simulate heterogeneous HIV samples of different diversities, evolved from a single parent genome[54]. Reads were generated from these populations by mixing the haplotypes in various proportions and coverage depths using Metasim, which replicates the error process of 454/Roche sequencing.

We require the size of l-tuple to be large enough to avoid repetition within the genome, yet be small enough that a large number of reads contain the l-tuple. We select a lower bound on l as $1/p^l > G$, where p is the probability that the most frequent nucleotide will appear at a given position and G is the genome size. When G is unknown, we approximate l using $p^l > NL$, where N is the number of reads and L is the average length of reads. Empirically, l-tuple length of 10 bp gives us optimal results. The read length of 250 bp was used for our experiments. Unless otherwise indicated, a default error rate of 0.5% was used for the simulations. Note that in our method, even a single base difference, if it appears consistently in all copies, will lead to its identification as a new haplotype. We have decoupled our analyses in the order of three tasks that we perform; error correction, greedy heuristic for predicting frequency estimation and clustering of l-tuples.

Figure 5.4 shows the snapshot of a histogram of l-tuples along the length of the genome for a viral population containing three haplotypes sampled in

mixing proportions of 1:3:5. Note that the l-tuples sampled from suspect regions appear with a low count relative to those sampled from the rest of the genome. By determining the number of times each l-tuple occurs and the fraction of the genome it accounts for, we can uniquely identify the tuples belonging to the suspect regions.

Our goal is to identify mutations that co-occur in each subset of haplotypes and the frequency with which they appear in the viral population. We therefore report *precision* and *recall* averaged over the different suspect regions, i.e. over the l-tuples that vary between the haplotypes and that do not include the large number of l-tuples common to all. Our method estimates cluster C_{y_i} to contain l-tuples that span the polymorphic sites that co-occur in haplotypes $x_k \in P_{y_i}$. The counts of these l-tuples converge to mode $y_i = \sum_{x_k \in P_{y_i}} x_k$. Subsequently, the subset of haplotypes that contribute to this mode, is recovered by the greedy heuristic as P_{y_i}. Let T_{y_i} denote the *true* cluster assignment of such l-tuples. We define *recall* and *precision* as follows:

$$\text{Recall} = \frac{|\text{l-tuples in } T_{y_i} \cap \text{l-tuples in } C_{y_i}|}{|\text{l-tuples in } T_{y_i}|}$$

$$\text{Precision} = \frac{|\text{l-tuples in } T_{y_i} \cap \text{l-tuples in } C_{y_i}|}{|l\text{-tuples in } C_{y_i}|}$$

In order to obtain global performance statistics, we define *F-measure* as the weighted harmonic mean of precision and recall across different suspect regions.

Our method derives its strength from high and uniform coverage of the data. However, sequencing biases due to both statistical (eg. CG bias) and biological effects (eg. mappability bias) can skew the coverage ratios by causing certain regions of the genome to be over-sampled or under-sampled. Elimination of these biases is imperative for avoiding spurious conclusions regarding the data. In the presence of a reference genome, the adverse effects of such sequencing biases can be mitigated by normalizing the nucleotide bias in the data[1].

Error rates with Roche GS20 system have been estimated as approximately 5 to 10 error per kbp[26, 69]. In order to test the robustness of the algorithm to the presence of sequencing errors, we simulated datasets with sequencing errors varying from 3-40 errors per kbp, using an error model based on *Illumina* sequencing

Table 5.1. Effect of error correction using spectral alignment with thresholding on reduction in the number of erroneous l-tuples. Parameters: Simulated datasets with two haplotypes in mixing proportion of 1:3, at diversities of 4-6%, containing 10,000 reads. If we consider an error rate of 0.5% per bp, then approximately 4.9% of the l-tuples and 71% of the reads are expected to be contaminated with at least one sequencing error.

Error rate(%)	F-measure (% Erroneous l-tuples)	
	No Error Correction	Thresholding with Spectral Alignment
0.4	98.6(2.8)	99.7(0.04)
0.8	84.1(4.4)	99.3(0.1)
1.4	27.8(6.6)	98.9(0.3)
1.9	34.9(11)	98.1(0.4)
2.6	28.3(14.6)	97.4(0.9)
2.9	28(17)	49.6(2.2)
3.4	27(18.8)	32.4(4.7)

technology. Sequencing errors produce an excess of l-tuples that appear only once in the population, as opposed to polymorphic sites that appear several times. The true distribution of l-tuple counts is expected to be a mixture of an exponential distribution, for erroneous l-tuples, and a series of Poisson distributions describing the true l-tuple counts[75]. If we consider an error rate of 0.5% per bp, an l-tuple length of 10 bp and read length of 250 bp, then approximately 4.9% of the l-tuples and 71% of the reads are expected to be contaminated with at least one sequencing error.

We combine two existing techniques to handle errors. Prior to clustering, our method makes use of *spectral alignment*, described in [47], to perform error correction. Subsequently, we configure our method to discard l-tuples that appear below a certain threshold. The first minimum in the frequency spectrum is defined as the cutoff threshold for errors. More tailored approaches for error correction have been discussed elsewhere[35, 73, 74, 75]. We simulated datasets consisting of two haplotypes in mixing proportion of 1:3, at diversities of 4-6%. We report F-measure combined across different suspect regions, weighted by the number of l-tuples in each suspect region (Table 1). We observe that the error correction scheme outlined above resulted in significant performance gain. The

frequencies estimated were highly correlated with the true mixing proportions. Thus, Mutant-Bin is fairly robust to error rates of 2%. However, when the error rate exceeds 2.5%, the accuracy drops sharply, since more than 22% of the l-tuples are expected to be contaminated with errors.

We assessed the ability of the algorithm to recover the true haplotype frequencies using the greedy heuristic. We consider populations consisting between 2 and 5 haplotypes, at diversities ranging from 0.1-10%. The greedy heuristic finds the minimal generating set and hence, most parsimonious set of haplotypes, such that the observed frequencies for the l-tuples are best explained by a combination of *basic frequencies* within the generating set. Figure 5.6 shows the haplotype frequencies as estimated by our method. In the same figure, we plot the frequency estimated by state-of-the-art ShoRAH[73] on the same dataset is shown for comparison. ShoRAH makes use of a reference genome to extract a minimal subset of haplotypes that explain all observed read, while Mutant-bin clusters the l-tuples by haplotypes with similar frequency in the populations. All parameters of ShoRAH were set to default values and the algorithm was run for 5000 iterations. The mixing proportions predicted by our method are comparable to that of ShoRAH. Note that our method being frequency-based will conflate haplotypes with identical frequencies. For instance, for a population containing four haplotypes at a ratio of 1:3:3:10, our method bins the l-tuples into three clusters, with predicted frequencies in the ratio 1:3:10 (Figure 5.6).

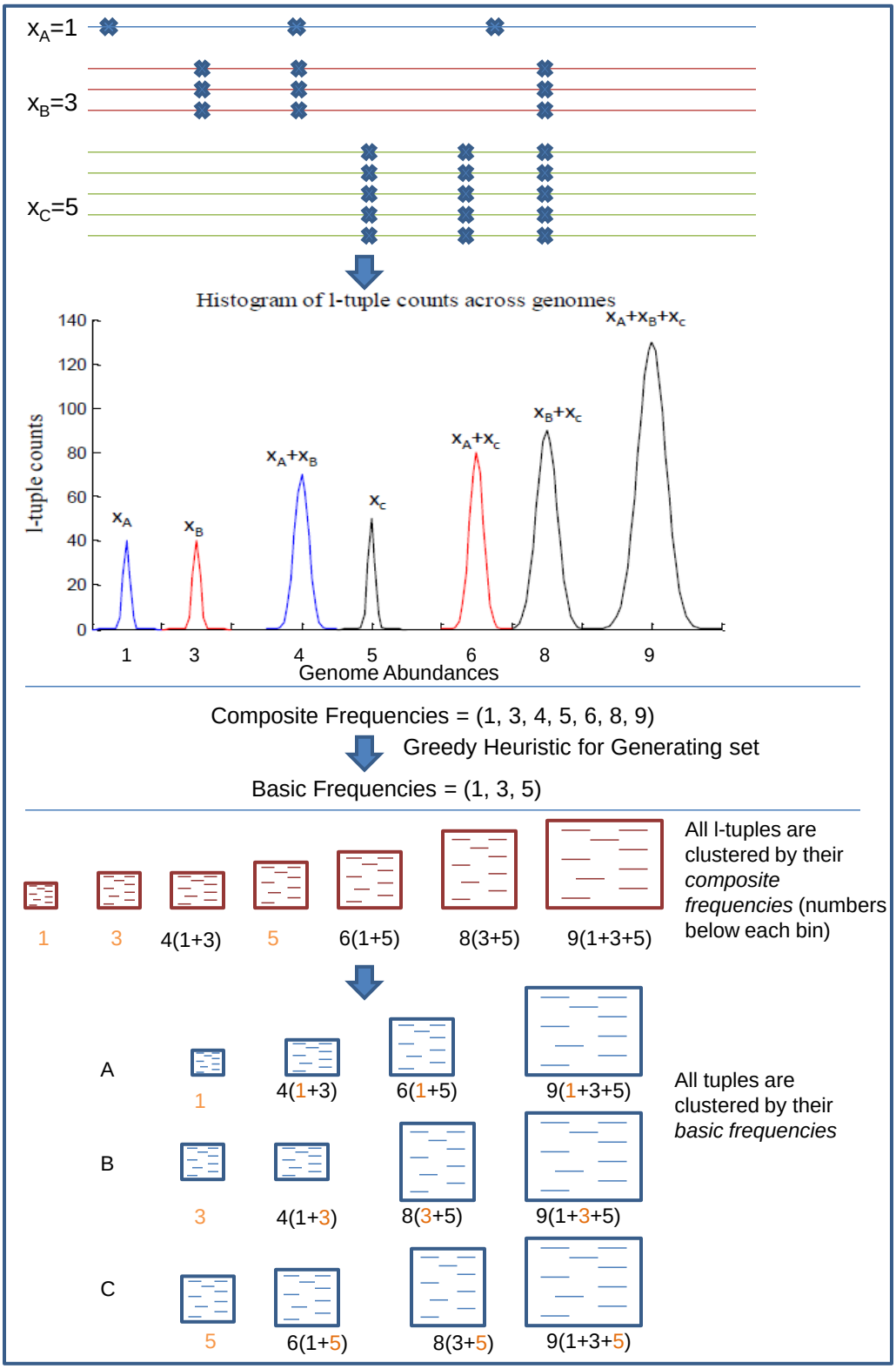


Figure 5.3. Diagramatic illustration of Mutant-Bin

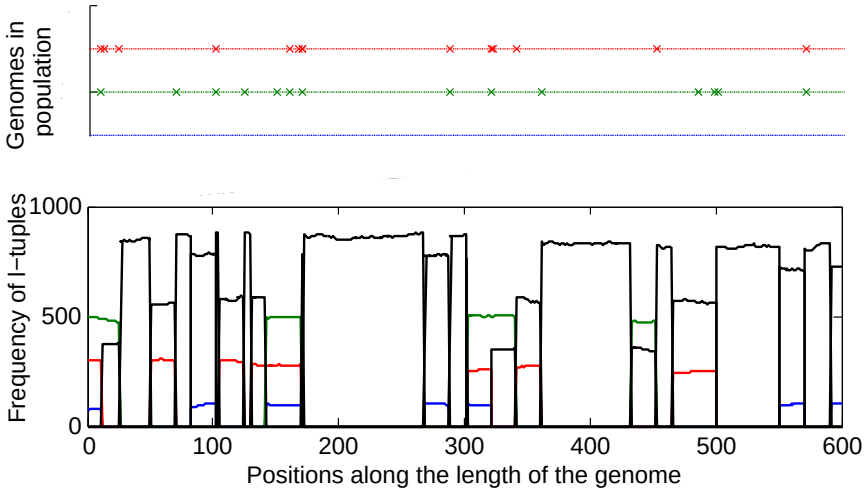


Figure 5.4. Top. Snapshot of 600 bp length of three genomes in a sample (with the crosses representing the derived nucleotides). The lowest genome is the designated reference. **Bottom**. Frequency spectrum of l-tuples count along the length of the genome for a viral population containing three haplotypes in the mixing ratio of 1:3:5.

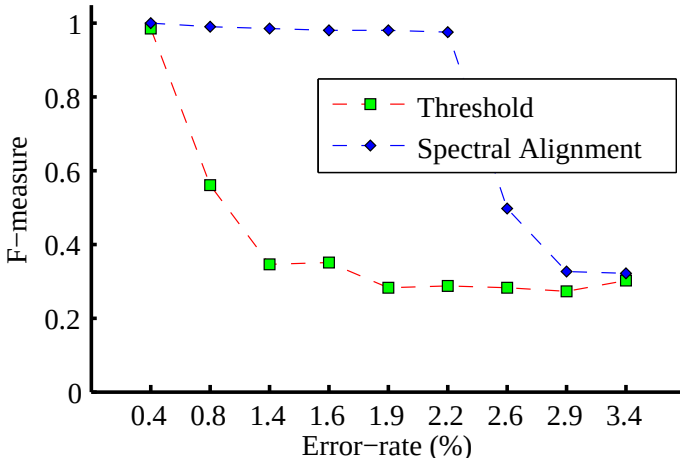


Figure 5.5. Comparison of effectives of two different error correction techniques 1. Thresholding of l-tuples 2. Spectral Alignment

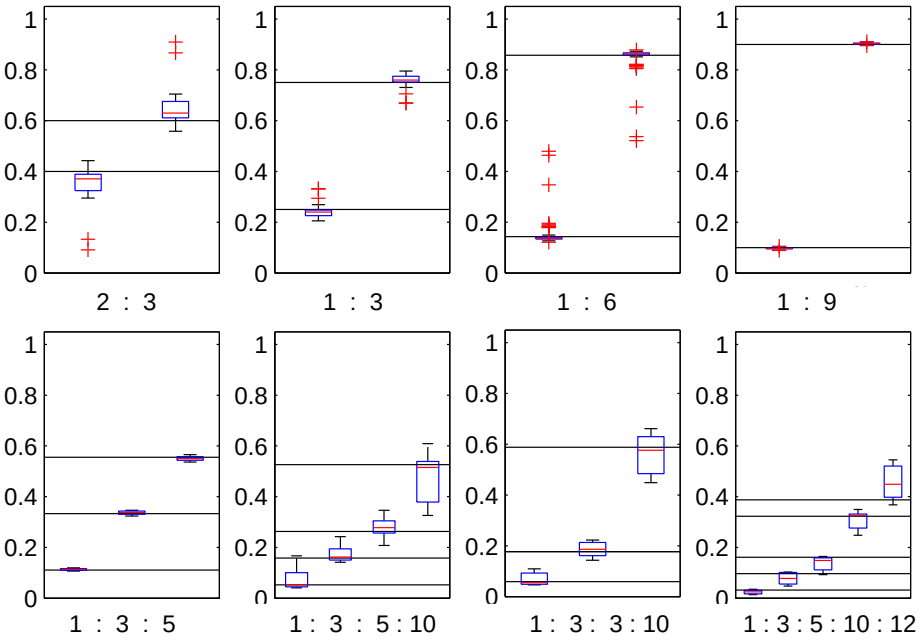


Figure 5.6. Estimated haplotype frequencies in sampling from mixing ratios (solid black lines) indicated beneath the panel for diversities 0.1-10%.

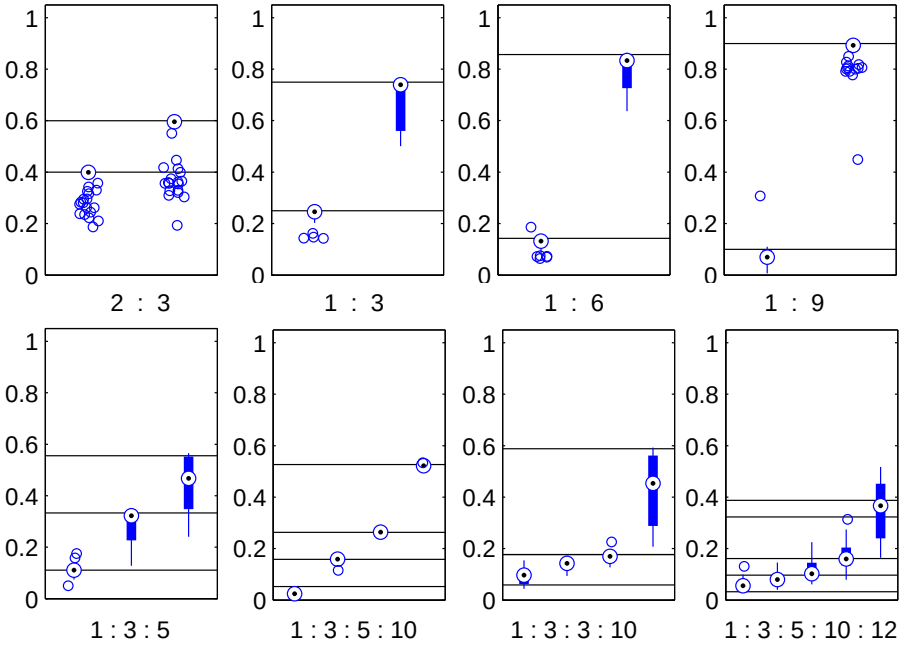


Figure 5.7. Estimated haplotype frequencies by ShoRAH in sampling from mixing ratios (solid black lines) indicated beneath the panel for diversities 0.1-10%. All parameters of ShoRAH were set to default values and the algorithm was run for 5000 iterations.

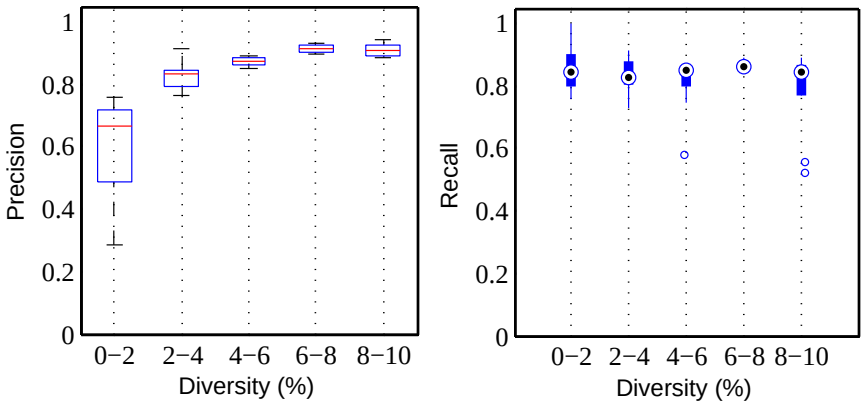


Figure 5.8. Precision**(Left)** and recall**(Right)** for datasets containing 2 haplotypes in the ratio shown to the left of the figures. X-axes shows the diversities varying from 0.1-10%.

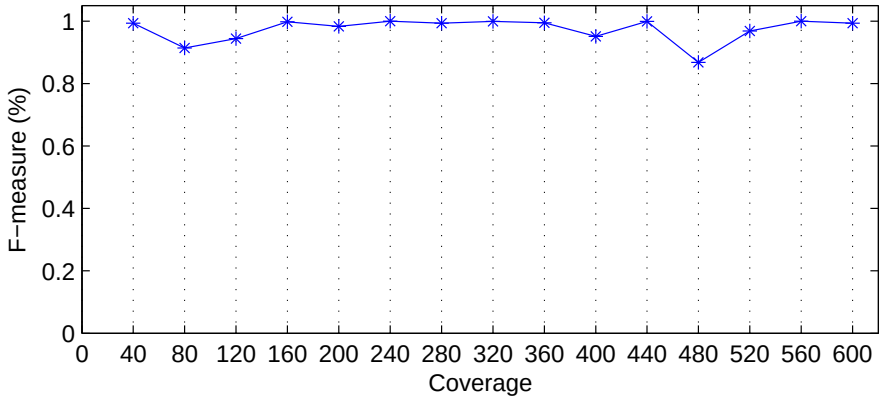


Figure 5.9. F-measure with varying coverage. Parameters:Simulated datasets with two haplotypes in mixing proportion of 1:3, at diversities of 4-6%

Next, we assess the sensitivity of the method to diversities. We consider populations chosen at diversities varying uniformly from $0.1 - 10\%$ (Figure 5.8). A total of 10,000 reads of average length 250 bp were drawn from sequences of length 2 kb. A high recall implies that we can identify with a high confidence, the set of l-tuples spanning the polymorphic sites that co-occur in a set of haplotypes. Figure 5.8 illustrates high recall values above 0.85 at virtually all diversities. Precision, on the other hand, depends strongly on the diversity. The main problem at low diversities is that sequencing errors can masquerade as polymorphic sites, resulting in a low precision.

The mixing proportions estimated by our method are in good agreement with the true mixing proportions. Since our method is a frequency-based method, we observe a higher deviation in the estimated fractions when the mixing proportion is close to 1:1 and with increase in number of haplotypes. The overall accuracy for populations with different number of haplotypes is shown in Figure 5.10. For five or more haplotypes, we obtain a high precision at the cost of low recall. Even when the cardinality of generating set obtained using the greedy heuristic is greater than the optimal, the true basic frequencies were correctly identified and formed a subset of the generating set. In no case did we find a generating set that is smaller than the set actually used to create the instance.

Our final evaluation on real deep sequencing HIV data obtained from 454/Roche FLX pyrosequencing platform allows for hard assessment of the performance[73].

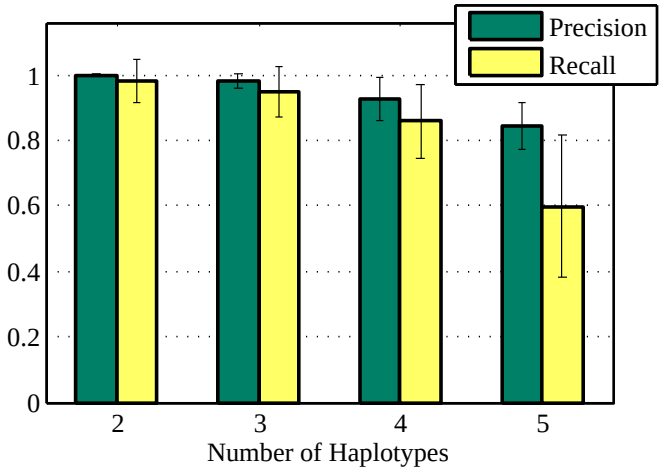


Figure 5.10. F-measure with the number of haplotypes in the population. Number of reads = 20000. Haplotypes were considered with frequencies corresponding to 1:3:5:10:12.

Of the four datasets, two are of subtype A and two of subtype B. The 454 reads from the two subtypes, which are at a diversity of 10.9%, were taken in the proportion of 1:4. After error correction by spectral alignment, our method obtained an F-measure of 87.3% averaged over the variable l-tuples from the suspect regions and accurately predicted the haplotype frequencies. For the same dataset, ShoRAH reconstructed haplotypes which are at a distance of less than 1% from the original haplotypes. Simultaneously, we generated reads *in-silico* with 454 sequencing errors from the 1.5 kbp region of HIV *pol* gene that has been sequenced into 4 clonal sequences[74]. We applied the method to read data mixed in proportions of 1:3:5:10, which are at a diversity of 7.8%. Our method achieved an F-measure of 94.2% and identified the frequencies of the three most frequent haplotypes correctly.

5.4 Conclusion

In this chapter, we have proposed an unsupervised method for quantifying the genetic diversity of heterogeneous populations in datasets for which no reference genome is available. Phylogenetic and evolutionary studies that do not require the knowledge of the sequence itself, but only the number of variable sites and the

nucleotides in these sites can benefit from such a method. A crucial advantage of our method is that it can reliably detect the co-occurrence of polymorphic sites in a subset of haplotypes and the frequency with which they appear in the population. It determines the population structure and haplotype frequencies without the use of a reference genome and hence, avoids the costly clonal sequencing and the compute intensive alignment required by other methods. It does not require a priori knowledge of the number of haplotypes. Note that our method does not reconstruct the haplotypes, it identifies suspect regions within the population. Deep sequencing technologies can produce much shorter reads of about 36 bp length, at a higher coverage. Our method is especially suitable for such short reads, as long the length of the reads exceeds the length of the l-tuple.

In its current implementation, our model possesses the following limitations. The method performs very well on datasets obtained with high coverages. However, in real sequencing projects, the frequency spectrum will be obscured by errors stemming from the sequencing process and non-uniform coverage across the genomes. Our method relies on the *basic frequencies* of the haplotypes in the population being distinct, which cannot always be guaranteed. Determining exact values of *composite frequencies* is the only step that is difficult with the increase in the number of haplotypes and reduced coverages. However, with better mode detection algorithms and high depth provided by NGS technologies, it should be possible to obtain a good separation between the distributions and improve the scalability of the method with the number of haplotypes. Mutant-Bin is a work in progress. Currently, the output of our method is a cluster of l-tuples, each corresponding to a haplotype in the population. We intend to extend our method to enable reconstructing the l-tuples in each cluster, into a genome of the corresponding haplotype. We also plan to incorporate the evolutionary tree-construction algorithm within the framework.

We foresee the application of this method in the context of cancer and bacterial communities that are also characterized by increased genetic diversity. The method can also be used for calling Single Nucleotide Polymorphisms (SNPs) (single nucleotide polymorphisms) without a reference genome such as for the endangered species[52]. A SNP commonly has only two haplotypes in the population and hence, determining the *composite frequencies* will be much easier.

Chapter 6

Conclusion

Genomics and Bioinformatics have the power to increase our understanding of all facets of life. The last decade has seen a massive explosion in the amount of biological information available due to huge advances in the fields of genomics and bioinformatics. The wealth of data has resulted in numerous discoveries that have had profound impacts on fields from vaccine discovery to agriculture, environment energy to anthropology. Genomics has deepened the need to understand the code of life, DNA. In recent times, large-scale DNA sequencing and assembly experiments have had a huge impact in the growth of the science of bioinformatics. The ultimate goal of bioinformatics and genomics is to develop supporting computational framework that will enable scientists to study the full diversity of the microbial world, their functions and evolution, in their natural environments[24].

The importance of metagenomics arises from the fact that over 99% of the species yet to be discovered are resistant to cultivation. Unlike single genome sequencing, assembly of a metagenome is intractable and is by large, an unsolved mystery. Moreover, the advent of high throughput sequencing is fueling rapid generation of enormous metagenomic datasets. We need to determine the number of species in a metagenomic dataset as well as the abundance of each of these species. In this thesis, I have addressed the challenge of analyzing and handling these datasets by implementing methods that enable scientists to study the full diversity of the microbial world including their functions and evolution in their natural environments.

6.1 Summary of Contributions

Clustering methods aim to identify the species present in the sample, classify the reads by their species of origin and quantify the abundance of each of these species. Clustering provides deeper insight into the structure of the community. The efficacy of clustering methods depends on the number of reads in the dataset, the read length and relative abundances of source genomes in the microbial community. In this context, in my dissertation research, I have presented methods to characterize and analyze the taxonomic content of vast amounts of short metagenome reads and heterogeneous viral populations.

In chapter 2, we proposed a two-pass semi-supervised algorithm for fuzzy clustering of metagenome reads that is a hybrid of comparative and composition based approaches. This method significantly reduces the size of the metagenome dataset while maintaining an accurate representation of its functional and taxonomic content. Overlapping clusters generated by a fuzzy clustering algorithm elegantly handle the problems associated with the nature of metagenomic data while providing tolerance for the noise in the data due to errors in sequencing and fragmentation. Our primary goal is to enrich the dataset into a small number of clusters such that reads within a cluster are phylogenetically closer than reads from different clusters. In our method, the comparative analysis of reads avails a priori biological knowledge in the existing database to form an initial set of seeded clusters. In the following pass, the composition based characterization of the remaining fraction of reads into existing clusters, facilitates a means of exploring novel species. The secondary goal is to identify polymorphic and conserved regions and capture them within the soft boundaries of the clusters. Due to evolution, the nucleotide composition of genomes belonging to same lower taxonomic levels can be very similar. Regions with overlapping clusters capture reads that are phylogenetically closer.

The significantly reduced size allows a compact yet comprehensive overview of the dataset. The proposed method does not require assembled contigs or training on a reference set, nor does it make any assumptions on the number of species or the nature of the dataset. It makes use of a reference database, however is not dependent on it. Our method enriches the dataset into a small

number of clusters, while accurately assigning fragments as small as ~ 100 base pairs. An important consequence of our method is that the fuzzy boundaries between clusters capture the misplacements of reads due to over representation of conserved regions, without clipping potentially useful sequences. In Chapter 2, we present the algorithm in detail and discuss the experimental results on two datasets: a simulated dataset of 454 reads that are 100 bps at varying coverage, and acid mine drainage metagenome dataset.

In Chapter 3, we formulated an unsupervised naive Bayes multi-species, multi-dimensional mixture model for reads from a metagenome. We use the proposed model to cluster metagenomic reads by their species of origin and to characterize the abundance of each species. Recent studies in metagenomics indicate the presence of a "genome signature", a compositional parameter which reflects the relative abundance of different words along a genome that can be used to distinguish between reads from different species. We model the distribution of word counts along a genome as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare. We employ either a mixture of Gaussians or mixture of Poissons to model reads within each bin. Further, we handled the high-dimensionality and sparsity associated with the data, by grouping the set of words comprising the reads, resulting in a two-way mixture model. We demonstrated the accuracy and applicability of this method on simulated and real metagenomes. Our method outperforms LikelyBin, another unsupervised composition-based binning method for metagenomes, on datasets of varying abundances, divergences and read lengths.

In Chapter 4 we presented an efficient multivariate Bayesian mixture model based on Poisson and Multinomial distributions for clustering discrete sequence data. The structure of Bayesian networks efficiently encodes the conditional dependencies between the words due to overlaps. The Poisson mixture model is derived from the assumption that the distribution of word counts along a genome follows a Poisson distribution. The Multinomial mixture model is derived as a standardized Poisson mixture model. We present a two-way clustering approach to handle the high-dimensionality and sparsity associated with the data. It combines the words along the reads into word groups and then constrains the parameters for words within the same group to be identical. The motivation of this

method is to overcome the bottleneck of Naive Bayes by taking into account the conditional dependencies between the word counts within the reads.

Our method can cluster reads as short as 50 bps with accuracy over 80% and estimate species abundance as well. The Bayesian mixture of Poissons and Multinomials outperform their Naive Bayes counterparts on datasets of varying abundances, divergences and read lengths. Our method is robust to the number of species in the dataset, read lengths and relative abundances of source genomes in the metagenome. Despite our specific application to metagenomics, the Bayesian mixture models are useful for classifying any high-dimensional discrete sequence data.

In chapter 5 we described a computational framework, called Mutant-Bin, for clustering individual haplotypes in a viral population and determining their prevalence, based on a set of deep sequencing reads. The main advantages of our method are that: (i) it enables determination of the population structure and haplotype frequencies when a reference genome is lacking; (ii) the method is unsupervised; the number of haplotypes does not have to be specified in advance; (iii) it identifies the polymorphic sites that co-occur in a subset of haplotypes and the frequency with which they appear in the viral population. The method was evaluated on simulated reads with sequencing errors and 454 pyrosequencing reads from HIV samples. Our method clustered a high percentage of haplotypes with low false positive rates, even at low genetic diversity.

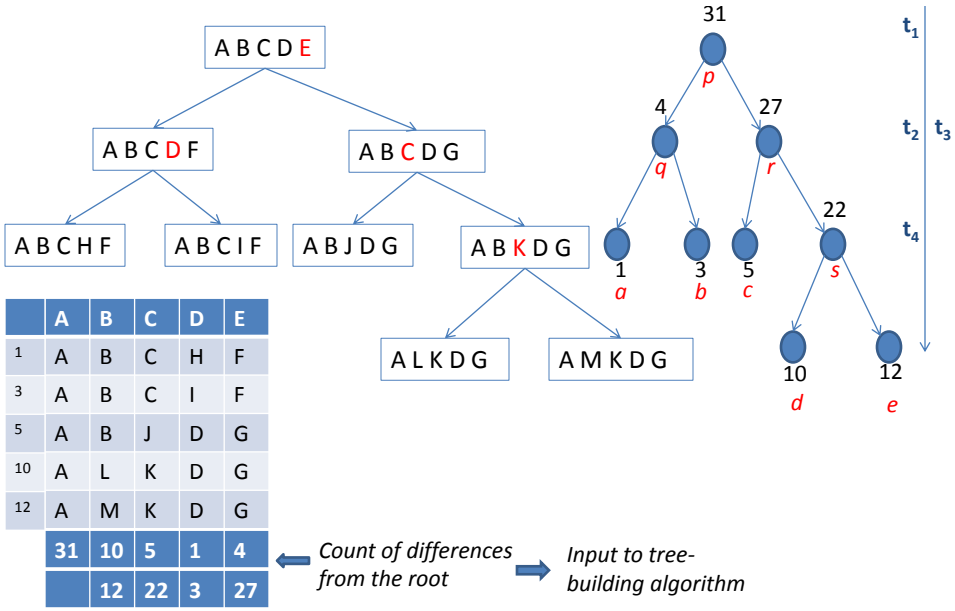
6.2 Future Work

Completely unsupervised clustering algorithm: The Naive Bayes and Bayesian clustering methods introduced by us in Chapters 3 and 4 are composition based methods. The performance of the proposed unsupervised composition based methods can be vastly improved by building a method that is a hybrid of composition based and similarity based methods, especially when evolutionarily close training genomes are available. It has been shown that unsupervised methods are effective on low complexity datasets, but less accurate on more complex datasets containing more than 20 different species.

All methods proposed until now require the user to supply the number of

species in each dataset as input. Determining the number of clusters from a statistical perspective is a difficult problem[65]. Maximum likelihood favors more complex models leading to over-fitting and hence is unable to address this issue. Previously, 16s/18s rDNA have been used for phylotyping and assessing species diversity using a rare-fraction curve. Most methods rely on heuristics to guide the choices of clusters. Determining species diversity is still an active area of research. A hierarchical approach to clustering that will clusters the reads in a metagenome at increasing taxonomic levels indicating the level of conservation fo the sequences at each level can be used to address the problem.

Inferring Phylogeny: Given a viral population of K haplotypes, the frequency spectrum represents the distribution of polymorphic sites with derived nucleotides that co-occur in a subset of haplotypes. Given how mutations occur and how a population evolves, it is possible to predict the shape of the spectrum in theory[28]. Using our framework for Mutant-Bin, we can work backwards to infer the population parameters and the evolutionary tree, given the shape of the spectrum. Such analyses play an important part in phylogenetics. The graphical structure of the tree is intrinsic to the co-occurrence of polymorphic sites within haplotypes. It is possible to derive the tree and branch lengths, given such a spectrumFigure 6.1. The polymorphic sites that have derived nucleotides, co-occurring in a subset of haplotypes $x_k \in P_{y_i}$ will appear with *composite frequency* y_i in the frequency spectrum. The number of such polymorphic sites shared between haplotypes in subset P_{y_i} will be proportional to the relative area of the Poisson curve with frequency y_i in the spectrum. Thus, one can infer the mutation probabilities $\mathbf{p}$ from the spectrum and then use $\mathbf{p}$ to infer the tree. Therefore, it is possible to construct the evolutionary tree without having to deal with the complications of the intermediate assembly step.



	A	B	C	D	E
1	A	B	C	H	F
3	A	B	C	I	F
5	A	B	J	D	G
10	A	L	K	D	G
12	A	M	K	D	G
	31	10	5	1	4
		12	22	3	27

Figure 6.1. Tree Construction: Using the frequency of differences of the viral genomes from the root of the tree, we can work backwards to reconstruct the evolutionary tree

Appendix

Definition 4 (Poisson distribution)**.** *If the expected number of occurrences of an event in a fixed interval is λ, then probability that there are exactly k occurrences is,*

$$P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Definition 5 (Multinomial Distribution)**.** *A multinomial distribution is a generalization of binomial distribution. In a bernoulli trial, there are two outcomes possible: success with a probability θ or $failure$ with a probability $(1-\theta)$. A binomial distribution is probability distribution of number of successes in n independent Bernoulli trials.*

$$P(X = k) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k} \text{ where } X = \text{ number of successes}$$

Similarly, in a catergorical trial, each trial results in a fixed number p of possible outcomes with probabilities $\theta_1, \theta_2, ..., \theta_p$, such that $\sum_{j=1}^{p} \theta_j = 1$ and each $\theta_j > 0, j \in 1, ..p$ and there are n independent trials. Let X_j indicate the number of times outcome j was observed over n trails. Then, the vector $(X_1, X_2, ..., X_p)$ follows a multinomial distribution with parameters n and Θ, where $\Theta = (\theta_1, \theta_2, ..., \theta_p)$. ∎ *That is,*

$$P(X_1 = x_1, X_2 = x_2, .., X_p = x_p) = \frac{n!}{x_1!x_2!..x_p!}\theta_1^{x_1}\theta_2^{x_2}...\theta_p^{x_p} \text{ where } \sum_{j=1}^{p} x_j = n \quad (1)$$

Theorem 6. *If $X_1, X_2, ..., X_p$ are independent Poisson variables with parameters,*

$\lambda_1, \lambda_2, ..., \lambda_p$ respectively (not necessarily equal), then the conditional distribution of $(X_1, X_2, ..., X_p)$ given that $X_1+X_2+...+X_p = n$ is multinomial with parameters λ_j/λ, where $\lambda = \sum \lambda_j$, i.e. $Mult(n, \pi)$, where $\pi = (\lambda_1/\lambda, \lambda_2/\lambda, ..., \lambda_p/\lambda)$.

In general, multinomial distributions are much easier to work with than Poisson, they both belong to the family of exponentials.

Proof. This proof is from STAT 504: Analysis of Discrete data(https://onlinecourses.science.psu.edu/s
Each X_j is distributed as:

$$P_{X_j}(x_j) = \frac{e^{-\lambda_j}\lambda_j^{x_j}}{x_j!}$$

The mutual independence of the X_j's shows that the joint probability distribution is given by:

$$P_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{p} \frac{e^{-\lambda_j}\lambda_j^{x_j}}{x_j!} = e^{-\lambda} \prod_{j=1}^{p} \frac{\lambda_j^{x_j}}{x_j!}$$

where $\mathbf{X} = (X_1, X_2, ..., X_p), \mathbf{x} = (x_1, x_2, .., x_p)$ and $\lambda = \lambda_1 + ... + \lambda_p$. Next, let $X = X_1 + X_2 + ... + X_p$. Then X is Poisson distributed with parameter λ (can be shown by using induction and mutual independences of X_j's):

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

The conditional probability distribution of X, given that $X = n$ is thus given by,

$$P_{\mathbf{X}}(\mathbf{x}|X = n) = \frac{P_{\mathbf{X}}(\mathbf{x})}{P_X(n)} = (e^{-\lambda} \prod_{j=1}^{p} \frac{\lambda_j^{x_j}}{x_j!})/(\frac{e^{-\lambda}\lambda^n}{n!}) = \frac{n!}{x_1!x_2!..x_p!} \prod_{j=1}^{p} (\frac{\lambda_j}{\lambda})^{x_j}$$

where $\sum x_j = n$ and $\sum \lambda_j/\lambda = 1$. □

This above theorem implies that the unconditional distribution of $(X_1, X_2, ..., X_p)$ can be factored into the product of two distributions: a Poisson for the overall total,

$$n \sim \phi(\lambda_1 + \lambda_2 + ... + \lambda_p)$$

and a multinomial distribution of X,

$$X \sim Mult(n, \pi)$$

The likelihood factors into two independent functions, one for $\sum_{j=1}^{p} \lambda_j$ and the other for π. The total n carries no information about π and vice-versa. Therefore, the likelihood based inferences about π are the same whether we regard $X_1, X_2, .., X_p$ as sampled from p independent Poissons or from a single multinomial. That is, any estimates, tests, etc. for π or functions of π will be the same whether we regard n as random or fixed. (**Important Point:** Our interest lies in the proportion of words in the reads. Inferences about the proportions will be the same whether we regard the sample size n as random or fixed. We will use this while formulating Bayesian mixture of Multinomials.)

Glossary

Assembly refers to the process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated. 1

BLAST (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. 7

Coverage is the number of times a genome has been sequenced. The Lander-Waterman equation for computing coverage is given by $C = NL/G$, where C stands for coverage, G is the haploid genome length. L is the read length and N is the number of reads. 3

DNA (Deoxyribonucleic acid) A nucleic acid that carries the genetic information in the cell and is capable of self-replication and synthesis of RNA. 1

Gene is a hereditary unit consisting of a sequence of DNA that occupies a specific location on a chromosome and determines a particular characteristic in an organism. 2

Genome is the total genetic content contained in a haploid set of chromosomes in eukaryotes, in a single chromosome in bacteria, or in the DNA or RNA of viruses. 1

Genus is a low-level taxonomic rank used in the biological classification of living and fossil organisms. 12

Haplotype is a set of polymorphisms or DNA variations that tend to be inherited together. 5

HIV (Human Immunodeficiency Virus) is a lentivirus that causes acquired immunodeficiency syndrome (AIDS), a condition in humans in which progressive failure of the immune system allows life-threatening opportunistic infections and cancers to thrive. 10

Metagenomics is the study of metagenomes, genetic material recovered directly from environmental samples. 3

Oligomers are words consisting of repeating units of nucleotides (A,C,T or G) and are of varying length. Specifically, k-mers or l-tuples are words of length k and l respectively. 14

Pathogen infectious agent is a biological agent that causes disease or illness to its host. 68

Phylum is a taxonomic rank below kingdom and above class. 13

Protein is the sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. 15

Proxygene is a full-length protein that has high local similarity with a specific short length of nucleic acid. 15

Quasispecies is a group of viruses related by a similar mutation or mutations, competing within a highly mutagenic environments. 68

RNA (Ribonucleic acid) is a nucleic acid present in all living cells. Its principal role is to act as a messenger carrying instructions from DNA for controlling the synthesis of proteins, although in some viruses RNA rather than DNA carries the genetic information. 26

Sequencing is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time. 1

Single Nucleotide Polymorphisms (SNPs) are DNA sequence variations occuring when a single nucleotide (A,C,T or G) in a genome differs (or is altered) between members of a biological species. 90

Bibliography

[1] Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads. *PLoS ONE*, 6:e16685+, 2011.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.

[3] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Măndoiu, P. Balfe, and A. Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC bioinformatics*, 12 (6), 2011.

[4] M. Bailly-Bechet, A. Danchin, M. Iqbal, M. Marsili, and M. Vergassola. Codon Usage Domains over Bacterial Chromosomes. *PLoS Comput Biol*, 2(4):e37+, April 2006.

[5] P. Beerli, N. Grassly, M. Kuhner, £. Nickle, O. Pybus, M. Rain, A. Rambaut, A. Rodrigo, and Y. Wang. Population Genetics of HIV: Parameter Estimation Using Genealogy-based Methods. In *Computational and Evolutionary Analysis of HIV Molecular Sequences*, chapter 10, pages 217–252. 2002.

[6] S. D. Bentley and J. Parkhill. Comparative genomic structure of prokaryotes. *Annual Review of Genetics*, 38(1):771–791, 2004.

[7] A. Brady and S. L. Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature methods*, 6(9):673–676, September 2009.

[8] V. Brendel, J. S. Beckmann, and E. N. Trifonov. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of biomolecular structure & dynamics*, 4(1):11, 1986.

[9] A. Campbell, J. Mrázek, and S. Karlin. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16):9184–9189, August 1999.

[10] C.-K. Chan, A. Hsu, S. Halgamuge, and S.-L. Tang. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, 9(1):215, 2008.

[11] S. Chatterji, I. Yamazaki, Z. Bai, and J. Eisen. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. *ArXiv e-prints*, 708, Aug 2007.

[12] K. Chen and L. Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol*, 1(2):e24, 07 2005.

[13] J. R. Cole, B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. The ribosomal database project (rdp-ii): sequences and tools for high-throughput rrna analysis. *Nucleic Acids Research*, 33(suppl 1):D294–D296, 2005.

[14] M. J. Collins, D. Kempe, J. Saia, and M. Young. Nonnegative integral subset representations of integer sets. *Inf. Process. Lett.*, 101:129–133, 2007.

[15] D. Comaniciu, P. Meer, and S. Member. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, 2002.

[16] D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *Proc. 8th Intl. Conf. on Computer Vision*, pages 438–445, 2001.

[17] D. Dalevi, N. N. Ivanova, K. Mavromatis, S. D. Hooper, E. Szeto, P. Hugenholtz, N. C. Kyrpides, and V. M. Markowitz. Annotation of metagenome short reads using proxygenes. *Bioinformatics*, 24(16):i7–i13, 2008.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38, 1977.

[19] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–1399, October 1999.

[20] N. Eriksson, L. Pachter, Y. Mitsuya, S.-Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS Comput Biol*, 4:e1000074, 2008.

[21] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 1968.

[22] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, J. Merrick, and e. al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.

[23] G. Folino, F. Gori, M. Jetten, and E. Marchiori. Clustering Metagenome Short Reads Using Weighted Proteins. In C. Pizzuti, M. Ritchie, and M. Giacobini, editors, *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, volume 5483 of *Lecture Notes in Computer Science*, chapter 14, pages 152–163. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2009.

[24] V. Hatzimanikatis and R. R. McCormick. Bioinformatics and functional genomics: challenges and opportunities. *AIChE Journal*, Vol. 46.

[25] C. Hoffmann, N. Minkah, J. Leipzig, G. Wang, M. Q. Arens, P. Tebas, and F. D. Bushman. Dna bar coding and pyrosequencing to identify rare hiv drug resistance mutations. *Nucleic Acids Research*, 35:91, 2007.

[26] S. Huse, J. Huber, H. Morrison, M. Sogin, and D. Welch. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8(7):R143+, July 2007.

[27] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–386, March 2007.

[28] P. L. Johnson and M. Slatkin. Inference of population genetic parameters in metagenomics: A clean look at messy data. *Genome Research*, 16:1320–1327, 2006.

[29] V. Jojic, T. Hertz, and N. Jojic. Population sequencing using short reads: Hiv as a case study. In *in Proc. Pac Symp Biocomput*, pages 114–125, 2008.

[30] J. Josse, A. D. Kaiser, and A. Kornberg. Enzymatic Synthesis of Deoxyribonucleic Acid. *Journal of Biological Chemistry*, 236(3):864–875, 1961.

[31] S. Karlin, I. Ladunga, and B. E. Blaisdell. Heterogeneity of genomes: measures and values. *Proceedings of the National Academy of Sciences of the United States of America*, 91(26):12837–12841, 1994.

[32] D. R. Kelley and S. L. Salzberg. Clustering metagenomic sequences with interpolated Markov models. *BMC bioinformatics*, 11(1):544+, Nov. 2010.

[33] A. Kislyuk, S. Bhatnagar, J. Dushoff, and J. S. Weitz. Unsupervised statistical clustering of environmental shotgun sequences. *BMC bioinformatics*, 10(1):316+, 2009.

[34] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231 – 239, 1988.

[35] H. Li, J. Ruan, and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18:1851–1858, 2008.

[36] J. Li and H. Zha. Two-way poisson mixture models for simultaneous document classification and word clustering. *Comput. Stat. Data Anal.*, 50:163–180, January 2006.

[37] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006.

[38] X. Li and M. S. Waterman. Estimating the Repeat Structure and Length of DNA Sequences Using -Tuples. *Genome Research*, 13(8):1916–1922, August 2003.

[39] B. Liu, T. Gibbons, M. Ghodsi, and M. Pop. Metaphyler: Taxonomic profiling for metagenomic sequences. In T. Park, S. K.-W. Tsui, L. Chen, M. K. Ng, L. Wong, and X. Hu, editors, *BIBM*, pages 95–100. IEEE Computer Society, 2010.

[40] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, In Press, Corrected Proof.

[41] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4(6):495–500, April 2007.

[42] A. C. C. McHardy, H. G. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods*, 4(1):63–72, January 2007.

[43] M. J. Morgan. A brief (if insular) history of the human genome project. *PLoS Biol*, 9(3):e1000601, 03 2011.

[44] S. Nasser, A. Breland, F. Harris, and M. Nicolescu. A fuzzy classifier to taxonomically group dna fragments within a metagenome. In *Fuzzy Information Processing Society, 2008. NAFIPS 2008. Annual Meeting of the North American*, pages 1 –6, May 2008.

[45] S. O'Neil, J. Dzurisin, R. Carmichael, N. Lobo, S. Emrich, and J. Hellmann. Population-level transcriptome sequencing of nonmodel organisms erynnis propertius and papilio zelicaon. *BMC Genomics*, 11:310, 2010.

[46] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, September 1988.

[47] P. A. Pevzner. A new approach to fragment assembly in dna sequencing. In *RECOMB*, pages 256–267, 2001.

[48] M. Pop and S. L. Salzberg. Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, 24(3):142 – 149, 2008.

[49] M. Prosperi, L. Prosperi, A. Bruselles, I. Abbate, G. Rozera, D. Vincenti, M. Solmone, M. Capobianchi, and G. Ulivi. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics*, 12:5, 2011.

[50] M. Qiao and J. Li. Two-way gaussian mixture models for high dimensional classification. *Statistical Analysis and Data Mining*, 3(4):259–271, 2010.

[51] M. S. Rapp and S. J. Giovannoni. The uncultured microbial majority. *Annual Review of Microbiology*, 57(1):369–394, 2003.

[52] A. Ratan, Y. Zhang, V. Hayes, S. Schuster, and W. Miller. Calling snps without a reference sequence. *BMC Bioinformatics*, 11:130, 2010.

[53] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and Statistical Properties of Words: An Overview. *Journal of Computational Biology*, 7(1-2):1–46, February 2000.

[54] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson. Metasima sequencing simulator for genomics and metagenomics. *PLoS ONE*, 3:3373, 2008.

[55] S. Robin, F. Rodolphe, and S. Schbath. *DNA, Words and Models: Statistics of Exceptional Words*. Cambridge University Press, November 2005.

[56] G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj. Metagenome fragment classification using n-mer frequency profiles.

[57] G. Russell, P. Walker, R. Elton, and J. Subak-Sharpe. Doublet frequency analysis of fractionated vertebrate nuclear dna. *Journal of Molecular Biology*, 108(1):1 – 20, 1976.

[58] M. N. M. S. Asharaf. An adaptive rough fuzzy single pass algorithm for clustering large data sets. In *Pattern Recognition*, 2003.

[59] F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen. Nucleotide sequence of bacteriophage [lambda] dna. *Journal of Molecular Biology*, 162(4):729 – 773, 1982.

[60] F. Schreiber, P. Gumrich, R. Daniel, and P. Meinicke. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, 26(7):960–961, Apr. 2010.

[61] B. W. Silverman. *Density estimation: for statistics and data analysis*. Chapman and Hall, 1986.

[62] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 515–522, New York, NY, USA, 2008. ACM.

[63] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6(9):938–947, September 2004.

[64] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. Glockner. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, 5(1):163+, October 2004.

[65] R. Tibshirani and G. Walther. Cluster Validation by Prediction Strength. *Journal of Computational & Graphical Statistics*, 14(3):511–528, September 2005.

[66] N. S. Toru Matsui, Gaku Tokuda. Termites as functional gene resources. *Recent Pat Biotechnol*, 3(3):1872–2083, 01 2009.

[67] B. A. Turlach. Bandwidth selection in kernel density estimation: A review. In *CORE and Institut de Statistique*, 1993.

[68] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, March 2004.

[69] C. Wang, Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R. W. Shafer. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome research*, 17(8):1195–1201, Aug. 2007.

[70] A. Willse and B. Tyler. Poisson and multinomial mixture models for multivariate sims image segmentation. *Analytical Chemistry*, 74(24):6314–6322, 2002.

[71] J. C. Wooley, A. Godzik, and I. Friedberg. A Primer on Metagenomics. *PLoS Comput Biol*, 6(2):e1000667+, Feb. 2010.

[72] Y.-W. Wu and Y. Ye. A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-Tuples. 6044:535–549, 2010.

[73] O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *Journal of computational biology*, 17:417–428, 2010.

[74] O. Zagordi, R. Klein, M. Dumer, and N. Beerenwinkel. Error correction of next-generation sequencing data and reliable estimation of hiv quasispecies. *Nucleic Acids Research*, 38:7400–7409, 2010.

[75] X. Zhao, L. E. Palmer, R. Bolanos, C. Mircean, D. Fasulo, and G. M. Wittenberg. Edar: an efficient error detection and removal algorithm for next generation sequencing data. *Journal of computational biology*, 17:1549–1560, 2010.

Vita

Shruthi Prabhakara

Shruthi Prabhakara is Ph.D. candidate in the Department of Computer Science and Engineering at Pennsylvania State University. Prior to joining PSU in Fall 2007, she received a Bachelors in Technology in Computer Science from National Institute of Technology Karnataka, Suratkal (formerly KREC) in 2007. She was a research intern at Microsoft, during the summer of 2008, during which she developed a handwriting recognizer for an Indian language Tamil for a Tablet PC which can recognize 156 classes of handwritten Tamil characters with an accuracy of 90.8%. In the summer of 2010, she was a research intern at The John Craig Venter Institute under the guidance of Dr. Natalie Fedorova. While at JCVI, she developed a clustering algorithm for determining the taxonomic composition of the termite gut metagenome.