Winter 12-18-2021

Differential Methods In Modern Biological Data Analysis

Micah Thornton
Southern Methodist University, mathornton@smu.edu

DIFFERENTIAL METHODS IN MODERN BIOLOGICAL DATA ANALYSIS

Approved by:

Dr. Monnie McGee
Associate Professor of Statistical Science

Dr. Daniel Heitjan
Professor of Statistical Science

Dr. Raanju Sundararajan
Assistant Professor of Statistical Science

Dr. Daehwan Kim
Assistant Professor of Bioinformatics

DIFFERENTIAL METHODS IN MODERN BIOLOGICAL DATA ANALYSIS

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Biostatistics

by

Micah A. Thornton

B.S., Statistical Science, Southern Methodist University
B.S., Computer Engineering, Southern Methodist University
M.S., Computer Engineering, Southern Methodist University

December 18, 2021

ACKNOWLEDGMENTS

I would like to thank the members of my dissertation committee: Drs. Heitjan, Kim, McGee, & Sundararajan for their time, patience, and valuable insights in the preparation of this manuscript. I would especially like to thank my advisors: Dr. Kim and Dr. McGee for their invaluable guidance in this work. The work in chapter three would not be possible without the UT Southwestern Lenette Lu Lab, thank you for your contributions and conversations.

Micah A. Thornton B.S., Statistical Science, Southern Methodist University
B.S., Computer Engineering, Southern Methodist University
M.S., Computer Engineering, Southern Methodist University

Differential Methods in Modern Biological Data Analysis

Advisor: Professor Monnie McGee
Doctor of Philosophy degree conferred December 18, 2021
Dissertation completed September 3, 2021

Analysis of biological data for differentiation of organisms/cells within and across species or even the same organism is important to a wide variety of applications. This work considers three different biological data sets at the genome, proteome, and epigenome levels: respectively, DNA sequences, glycosalation data, and DNA methylation. We explore some statistical modeling approaches for handling these modern datasets, and provide a relevant set of experiments for explanation and illustration.

First, genomic Fourier coefficients, which capture information about the harmonics of genetic sequences in terms of nucleotide pattern recurrence are investigated as summary metrics for medium sized virus genomes from the SARS-CoV-2 virus. Clustering and classification techniques are applied to these for identification of the geographic location of submission of the original sample. It is shown that the Fourier coefficients are potential features on which geographic location of sequences can be classified with 79% accuracy. Furthermore, the Fourier coefficients provide distance metrics for efficient clustering.

Second, at the protein expression level, we describe data that measure the composition of protein glycosalation in tuberculous patients and perform studies to use the glycosalation profiles as markers for patients with a particular disease status. Three models are discussed: a classical approach known as partial least squares discriminant analysis (PLS-DA), and two new approaches which are developed for general datasets with compositional data. These models are examined using protein data from capillary electrophoresis (CE) quantification

of glycan species in tuberculosis patients. The models show a marginal improvement over the PLS-DA approach, 45% accuracy over 41% (five-fold cross validation, with five outcome categories).

Third, at the epigenetic level, we discuss a critique of the use of local likelihood regression smoothing to determine methylation via Bisulfite sequencing. We show a relationship between the sensitivity of these windowed averaging techniques and variations in the coverage of methylated areas via simulation. A procedure for combining read densities with methylation information to resolve multi-mapped reads is described.

TABLE OF CONTENTS

LIST OF FIGURES **x**

CHAPTER

1. Introduction **1**

1.1. Modeling the Harmonics in SARS-CoV-2 Genomes 2

1.1.1. Fourier Transform Background 3

1.2. Modeling the Variation in Glycosylation on Antibodies 5

1.3. Modeling the Methylation Profile of a Genetic Sequence 7

2. Analysis of the Fourier Coefficients of Genomic Sequences **9**

2.1. Fourier compared to Walsh Transform 11

2.1.1. Encoding & Transforming Genomic Sequences 12

2.2. Data Overview 16

2.3. Methods and Procedures 18

2.3.1. Comparison to Other Distance Metrics 24

2.3.2. Filtering Procedures 26

2.3.3. Complex Gaussian Model for Fourier Coefficients 30

2.4. Results and Analysis 34

2.4.1. Clustering and Visualization for SARS-CoV-2 Sequences 34

2.4.2. Multivariate Analysis of Variance 38

2.4.3. Supervised Learning by Geographic Region 40

2.4.4. Correlation between Distance Methods 42

2.4.5. PS Filtering Results 43

2.5. Conclusions 48

3. Analysis of Compositional Data in Antibody Glycosylation **50**

3.1. Glycosylation on Antibodies in Disease Response 53

3.2. Data Description 54
3.3. Modeling Approaches 56
3.4. Partial Least Squares Modeling 57
3.4.1. Data Study: 2021 Tuberculosis Data - PLSDA 62
3.5. Semi-Parametric Modeling 64
3.5.1. Model Specification 66
3.5.2. Data Study: 2021 Tuberculosis Data - Semiparametric 72
3.6. Glycan rank probability method 76
3.6.1. Link of glycan rank probability to enzyme concentration 77
3.6.2. Model Specification 79
3.6.3. Model Application Results 80
3.7. Conclusions 82
3.7.1. Future Directions 84

4. Analysis of Epigenetic Signals of Genomic Sequences 85
4.1. DNA Methylation 86
4.2. Some Previous Statistical Approaches to Methylation 87
4.2.1. CpG point estimates of methylation 88
4.3. Potential Methylation Locations along A Reference (GRCh 38) 91
4.4. Reduction of Multi-Mapped Reads 95
4.5. Simulation of Methylation Profile Errors 98
4.6. Conclusions 100

5. Conclusion & Future Directions 102
5.1. Future Directions 103

APPENDIX

A. Genomic Fourier Coefficients: Tables & Figures 105

B. Modeling Compositional Data: Tables & Figures 108

REFERENCES **125**

LIST OF FIGURES

Figure		Page
2.1	Simulation Error Walsh V. Fourier Coefficients for size from 16 to 262,144	11
2.2	Flow Diagram of Genomic Fourier Analysis	15
2.3	Schematic of Machine Learning approach for automatic filter learning	28
2.4	Complete Fourier Spectra Visualization: TSNE plot	35
2.5	TSNE plots for different distance procedures	37
2.6	Fourier Coefficients Canonical Correlations Plot	40
2.7	Correlation Matrix for Distance Methods	42
2.8	Variance Diagnostics for Power Spectra	44
2.9	Some representative filters produced by the AFL method	45
2.10	Scree Plot and Bi-Plot	46
2.11	Filtering Methods Comparisons, by Correlation to Full PS Distances	47
3.1	Structure of IgG for reference	51
3.2	Example Data-Collection Multiple-Composition Creation	54
3.3	Types of Glycan Structures	55
4.1	KDE of CpG Island Distribution on Human Chromosome 22 (GRCh38)	91
4.2	Number of reads covering CpGs in human Chromosome 22	92
4.3	Variance in number of CpG "Hits" per Read Length	94
4.4	Average Error in Methylation Point Estimates over Varying Read Lengths from simulated ground truth	95
4.5	Average Error for windowed average smoothing approach on Simulated reads with synthetic ground methylation data	96
4.6	Illustration of the Multi-Mapped Read location isolation procedure	97
B.1	Plots of non-parametric variables within five classes	113

B.2 Partial Least Squares Tuberculosis/Diabetes Classification LV plot 115
B.3 Partial Least Squares Tuberculosis Classification LV plot 116
B.4 Partial Least Squares Diabetes Classification LV plot 117
B.5 Principal Components Plots Tuberculosis/Diabetes Biplot 118
B.6 TSNE Plots for Tuberculosis/Diabetes classification) 119
B.7 Class Accuracy for PLS-DA classifier model 120

To my parents Mitch & Misty, and brother Henry with much love.

Chapter 1

Introduction

The techniques discussed in this dissertation provide a set of potential approaches to handle specific types of data, where the distinction between certain classes is of interest. Specifically, classification using DNA sequences, glycosylation data, and methylation data are examined. In Chapter 2, the use of the Fourier transform of a numerically encoded version of a DNA sequence is discussed. The Fourier transform is used to locate periodic structure in the data, which can be considered as a type of signature for a given data set. Analysis of the coefficients produced by the Fourier transform, and the average coefficient magnitudes, power spectra (PS), as a numerical summary for classification of the sequences is investigated. Three post calculation filtering techniques for selecting important subsets and combinations of the variables are described. A hypothesis test for testing the null hypothesis that the variance-covariance matrices of complex Fourier coefficients among signals in different groups are all the same is derived. The techniques are applied to a dataset of SARS-CoV-2 genomes extracted from various geographic regions in order to classify the sequences by their region of origin.

In Chapter 3, three different techniques are used to model and quantify the glycan structure of immunoglobulin in patients with tuberculosis with or without diabetes mellitus. First, a partial least squares procedure is discussed and the results of its application in modeling the tuberculosis data are interpreted. Following this, A semi-parametric classification model for the data derived by treating compositional data in the model as observations of a multinomially distributed random variable is given. The multinomial model is combined with class estimates based on nonparametric functional data profiles. Finally, an approach involving estimation of one-step glycan transition probabilities within each class based on relative frequency of each class is discussed.

In the final chapter, we give a short explanation and introduction to the recent method of bisulfite-sequencing for mapping of DNA methylation regions to the genome. It is often the case that multiple reads are mapped to the same place on the genome; therefore, a method of determining the most likely position for methylation is introduced.

1.1. Modeling the Harmonics in SARS-CoV-2 Genomes

First, this dissertation is concerned with the repetitions and patterns of base nucleotides in a sequence, and describes the use of Fourier analysis for quantifying periodic behavior within an encoded genomic sequence. Fourier coefficients for a discrete time signal are given by Equation 1.1 [Brillinger, 2001].

$$\boldsymbol{d}_X^{(T)}(\lambda) = \sum_{t=0}^{T-1} X(t) e^{-i\lambda t} \tag{1.1}$$

Here t corresponds to the discrete time point at which $X(t)$, the signal, is observed. The overall signal length is denoted by T, the frequency of interest is λ, and e and i are used to represent the natural and imaginary numbers respectively.

Fourier analysis has been previously applied to DNA sequence data for determining phylogenetic relationships [Yin et al., 2014, Yin and Yau, 2015, Yin, 2020] and patterns with a sequence itself for alignment [Katoh et al., 2002], and detection of coding regions [Zhou et al., 2007]. It is clear that if two genomes are sequenced and found to be identical then *the sequences themselves* contain no humanly discernable information about the virus from which they were captured. That is to say, if two genomes were identical in sequence, and one were picked at random, then there would be no possible way of knowing whether the sequence selected originated from one organism or the other based on the sequence information itself. However, the periodic information of the sequence tends to change quickly, which could allow fast distinction of mutations across sequences.

The Fourier Transform is used to capture periodic information in numerical coefficients of genomic sequences. Those sequences are then compared through the numerical distances between those coefficients. One key result is a demonstration of this procedure using a viral data set curated by GISAID containing SARS-CoV-2 viral genomes. The use of filters on the coefficient values of the PS, and a hypothesis test for testing the null hypothesis of equivalent variance-covariance of the Fourier coefficients is also discussed.

1.1.1. Fourier Transform Background

The history of the Fourier Transform dates to the eighteenth century
[Heideman et al., 1984]. Trigonometric transforms were used by Euler and later formalized into the Fourier transform, which was finalized and popularized by Gauss. In its earliest incarnation, the transform was used to decompose continuous time functions into a constituent series of weights on the Fourier basis functions with increasing frequency. More recently the transform has been adapted and applied to analysis of discrete time signals, as well as otherwise indexed signals that show periodic trends. Two such examples are analysis of seasonal meteorological data [Salcedo and Baldasano Recio, 1984] and a study on measles outbreaks in England and Wales across seven separate epidemic outbreaks [Grenfell et al., 2001]. In the measles outbreak analysis, the DFT showed merit in the prediction of disease occurrence at specific locations, and at specific times. The transform has also found application in pharmaceutical discovery and validation; mainly in the context of discovery of various potentially therapeutic molecules [Song et al., 2020]. It has also been applied to the classification of epileptic vs. non-epileptic patients via signals from patient electroencephalogram (EEG) readings [Polat and Güneş, 2007]. For a general review of other applications of the Fourier and other transforms to medical sciences see the review chapter published in [Armăşelu, 2017].

In terms of genetics and genomic analysis, as well as bioinformatics specifically, the Fourier transform has had a wide variety of uses as well. One early application of the

DFT to genetic sequences used the periodic nature of codons (consisting of three consecutive nucleotides encoding amino acids in a protein) in order to differentiate coding regions of a genomic sequence from the non-coding regions [Fuentes et al., 2006]. Another popular application of the Fourier transform is found in genetic sequence alignment using a sliding window based matching algorithm, known as MAFFT (Multiple Alignment using Fast Fourier Transform) [Katoh et al., 2002]. In previous work it was shown that Kolmogorov-Smirnov statistics calculated on the distributions of Fourier and Laplace coefficients computed on the k-mer distributions of genetic sequences produced similar clusterings of the data [Thornton, 2019]. In addition, the Fourier transform has been used on DNA and protein sequences in order ascertain the existence and characteristics of periodic behavior in a genome [Yin et al., 2014, Yin and Yau, 2015, Yin, 2020]. Generally, nucleotide or protein sequences are encoded into binary or categorical value digit vectors prior to applying the transform. Encoding categorical series into discrete numerical signals is a non-arbitrary way to assign numerical values to characters used to describe DNA and protein sequences.

The Fourier transform may be thought of as a convolution of a signal with sine waves of consecutively increasing frequency. Mathematically, [Oppenheim et al., 1997], the transform of a continuous time signal can be written as:

$$a_n = \frac{1}{T}\int_0^{t=T} x(t)e^{-jn\omega_0 t}dt = \frac{1}{T}\int_0^{t=T} x(t)e^{-jn(\frac{2\pi}{T})t}dt \quad \text{If n} > 0$$

$$a_0 = \frac{1}{T}\int_0^{t=T} x(t)dt$$

Where the value T represents the total number of time points (or nucleotides in the sequence). $x(t)$ Represents the value at time t, in the case of genomic data $x(t)$ is a set of indicator functions indicating whether the sequence at loci t is a certain given nucleotide or not. j represents the imaginary number, and n is the desired coefficient of the Fourier spectra (which also modifies the frequency of the sinusoids with which the original function is convolved).

The use of spectral transforms in signal processing has seen broad general use in signals of many different kinds. These transforms allow their users to examine the signal in the frequency domain, rather than the initial domain. Examination of a signal in the frequency domain allows for determination of the periodic structure in the sequence, which can be used for unique identification of a sequence and clustering of similar sequences. Further, Fourier coefficients can be used to compress the original sequence without great loss of information. Both of these avenues are explored in Chapter 2.

1.2. Modeling the Variation in Glycosylation on Antibodies

Chapter 3 looks at analyzing protein compositions that may not be detected when relying on genomic or epigenomic data, particularly in relation to the differentation of disease subtypes by glycosylation of antibodies. A Partial Least Squares Regression (PLSR) [Geladi and Kowalski, 1986] approach known Partial Least Squares Discriminant Analysis (PLS-DA) [Pérez-Enciso and Tenenhaus, 2003],[Brereton and Lloyd, 2014],
[Gromski et al., 2015] is utilized to classify tuberculosis patients as active or latent according to their glycosylation profiles. In previous work, researchers applied the PLS-DA procedure coupled with cross validation to make determinations on subsets of glycosylation patterns to classify tuberculosis patients as active or latent. [Lu et al., 2020].

A description of the methodological approach for applying PLSR to general experimental data captured by labs is given in [Brereton and Lloyd, 2014]. PLSR can be extended to cases where there are more than two classes. PLS-DA procedure has been called 'an algorithm full of dangers' and the extension of the technique to multiple categories of has certain pitfalls and implications for the analysis. In Chapter 3, the use of PLS-DA on percentage of glycan species present in subjects is used to differentiate among five groups of patients. The groups of patients are active tuberculous patients with and without diabetes, latent tuberculous patients with and without diabetes, and patients negative for both tucerculous and diabetes.

Capillary Electrophoresis (CE) is the experimental procedure utilized to gain information about the glycan species which are present in the data [Hirabayashi, 2008],
[Mittermayr et al., 2013]. In CE, sampled glycans are placed at the entrance to a capillary tube upon which an electromagnetic field is induced [Gordon et al., 1988],
[Wallingford and Ewing, 1989], [Ewing et al., 1989], [Grossman and Colburn, 2012]. A detection plate sits at the other end of the capillary and is connected to a sensitive timer. This allows the recording of an intensity by time chart. From this chart, the intensities at specified time points, referred to as "peaks", allow the discernment of the relative concentrations of glycan species in the sample [Lu et al., 2018]. For a listing of the glycans and their respective structures, see Figure 3.3) The glycan sample contains a total fixed number of glycans of different species are well modeled using traditional compositional data techniques, as has been suggested and demonstrated in [Uh et al., 2020]. Studies using glycan profile analysis through mass spectrometry to quantify glycan species concentrations for the detection of arthritis [Nakagawa et al., 2007], lung adenocarcinoma [Lattová et al., 2020], and other cancers [Ruhaak et al., 2013], [Gebrehiwot et al., 2018], [Mechref et al., 2012] have shown success.

The term "compositional" refers to the nature of the data being split into multiple parts of a whole. There has been a long history of theoretical approaches applied to compositional data [Aitchison, 1982], [Pawlowsky-Glahn and Egozcue, 2006],
[Pawlowsky-Glahn and Buccianti, 2011], [Pawlowsky-Glahn et al., 2015], and
[Filzmoser et al., 2018b]. A wide variety of software has been produced for the analysis of the same [Van den Boogaart and Tolosana-Delgado, 2013]. Approaches that have been modified for accommodating the structure of these data include PCA [Aitchison, 1983], [Aitchison and Greenacre, 2002], general modeling procedures [Egozcue et al., 2003], and more [Aitchison, 1994]. In addition to PLSDA for classification of disease states using glycans, a semi-parametric model and a glycan rank probability model for the analysis of the same data are introduced.

1.3. Modeling the Methylation Profile of a Genetic Sequence

The sequential arrangement of the nucleotides in DNA is transcribed into RNA. The RNA sequence is vitally important to the composition and function of the proteins that are translated from it. However, secondary cues for the cell are encoded in chemical structures attached at vital points along the genomic sequence indicating the repression or enhancement of particular transcribed regions [Alberts et al., 2002] [Bonasio et al., 2010]. These referential reading guides on the genome are critically important for the cell, but may also be used to differentiate particular regions of interest among organisms, including in neurological disorders [Dall'Aglio et al., 2018], breast cancer [Liang et al., 2021], hypoxia and chronic ischemia [Eddy et al., 2019], and other observable phenotypes even differential migration patterns in fish [Baerwald et al., 2016]. DNA methylation, a type of epigenetic signal [Tate and Bird, 1993] [Edwards et al., 2017], and it's relationship with transcription binding factors [Yin et al., 2017] has also been investigated in cancer detection [Watanabe and Maekawa, 2010], adaptation to environment [Flores et al., 2013], and DNA repair [Walsh and Xu, 2006]. The methylation pattern of the DNA for a particular cluster of genes may exhibit statistically significant differences among two samples of the same genomic region [Stockwell et al., 2014], [Xie et al., 2019], [Park and Wu, 2016], [Chen et al., 2017]. These regions of differential methylation on the same genome known as Differentially Methylated Regions (DMRs) may be used to determine whether two organisms will express the same protein in the same place at the same time, for example, in disease progression.

One way in which the epigenetic cue of methylation is discovered in a sample of reads is via a procedure known as bisulfite sequencing [Darst et al., 2010], [Krueger et al., 2012]. It works by first treating the genetic material with bisulfite, a chemical which will convert unmethylated cytosine to thymine, while leaving the methylated variant untouched [Frommer et al., 1992]. Hence, when aligning to and comparing with a reference, the researcher is able to infer which reads (small captured fragments of roughly the same size captured in an experiment) indicate methylation or not for each particular location of inter-

est along the reference.

This dissertation does not seek to provide an exhaustive list of all of the procedures available for differentiating biological data, because such an undertaking would fill many volumes. Instead the primary focus of this dissertation work is on three specific areas of biological data: patterns in sequencing, compositions in glycosylation, and mapping of methylated reads in bisulfite sequencing. The procedures discussed in this dissertation represent important statistical considerations for some extant modeling procedures, and provide descriptions and examples of new ones. Specifically, novel investigations into subsequent modeling techniques using harmonic patterns in genetic sequences are provided, and a general code base in R [Thornton, 2021] for performing such an analysis is developed. Semi-parametric, and rank-based non-parametric methods for investigation compositional patterns in glycans are supplied. The modeling of methylation profiles with smoothing techniques such as BSmooth [Hansen et al., 2012], a local-likelihood procedure, are discussed, and their performance is compared against simulation displaying characteristics of simpler moving-averages. The moving averages have been shown to be almost equally effective as the more complex BSmooth [Wu et al., 2015]). Finally, an algorithm for refining an initial bisulfite sequencing alignment is provided.

The majority of the programming in this work was implemented in R [R Core Team, 2021], C [Kernighan and Ritchie, 2006], and MatLab [MATLAB, 2020]. In R the bioconductor [Morgan, 2021] suite was utilized for investigating the BSmooth procedure
[Hansen et al., 2012].

Chapter 2

Analysis of the Fourier Coefficients of Genomic Sequences

This chapter explores Fourier coefficients of genomic sequences, building in a few new directions on a procedure which has been adapted and implemented to various uses in sequence analysis since the 1990's [Anastassiou, 2001]. First, a sample of genetic transcripts sequenced from the SARS-CoV-2 virus are used to refine an R pipeline that calculates a Fourier transform for genetic sequences. A sample of 1,400 viruses are then used for the two primary experiments of this chapter [Elbe and Buckland-Merrett, 2017]. First, we show that the power spectra (PS) of the Fourier series for a genomic sequence can be used to classify the sequences by geographic region with approximately 80% accuracy. It is then shown that the PS can be used to provide distance metrics that augment extant metrics while not requiring sequence alignment.

The use of strategically selected (filtered) power spectral coefficients for the clustering and classification of sequences is also investigated. An analysis using filtered data shows that a very small subset of the overall coefficients will capture the information contained in the full set. This chapter also introduces and derives a statistical test of hypothesis for the Fourier coefficients, under the assumption that the sequences are second-order stationary. Finally concluding remarks and some future directions are suggested.

The motivating work for this research comes from a demonstration that the Fourier Transform of a genetic sequence provides useful phylogenies of organisms [Yin et al., 2014]. Yin, *et.al.* utilized Fourier transforms of NADH Dehydrogenase subunit-4 gene from the mitochondrial DNA of several primate species to produce a phylogenetic tree of the species. They compared their phylogeny to those formed from the standard Jukes-Cantor and K-mer distance approaches. They further investigated the ability of the procedure to correctly identify the phylogeny of sequences which were synthetically mutated. They showed that

the Fourier coefficients were able to produce phylogenies which were at least as accurate as those which were produced using the standard Jukes-Cantor procedure. Later, a more computationally efficient approach using Fourier transforms produced similar phylogenies [Yin and Yau, 2015], and is available in a suite of Matlab procedures [Yin, 2020].

The motivating question of this study concerns whether the Fourier coefficients alone are appropriate for unique identification of a sequence. From a theoretical perspective, neither the Fourier coefficients of a sequence, nor any other metric, can be used as a means to uniquely identify a sequences. For example, within the sample of 1,397 viral sequences which are included in this study, there are exactly 339 groups of sequences of which members are identical to one another, and hence cannot be differentiated by the computed Fourier coefficients. From the definition of the Fourier transform coefficients 1.1, it can be noted that even a single change will produce a different set of coefficients. One manner in which identical sequences might be differentiated in practice is by the addition of very small random errors to sequence coefficients, such that the relational structure of the sequences induced by the Fourier coefficients is maintained, while each sequence is supplied with a unique set of coefficients. This is also shown in previous work, when a viral sequence is artificially perturbed in demonstration of phylogeny building capabilities [Yin and Yau, 2015].

After the Fourier coefficients were determined for each of the sequences, a preliminary analysis seeking to determine how well the values cluster the data is performed. First T-Stochastic Neighbor Embedding (TSNE) [van der Maaten and Hinton, 2008] was applied to the Fourier coefficients, followed by Multivariate Analysis of Variance (MANOVA) [Wilks, 1932], and canonical variables analysis [Hotelling, 1935] [Kettenring, 1971]. The principal components of the coefficients were also considered, and finally several different classification procedures which are enumerated in the supplemental table were undertaken, and their accuracy assessed.

2.1. Fourier compared to Walsh Transform

Another of the theoretical questions to discuss when considering the Fourier Coefficients of a genomic sequence is whether applying the continuous domain transform to a discrete domain signal will produce results summarizing the frequency content that are too biased for any suitable use. That is, a genomic or genetic sequence, when encoded numerically, does not exhibit smooth transitions from one level to another, but rather instantaneous jumps. This indicates that it is more appropriate to model the frequency decomposition using square waves. The most natural transform for allowing this decomposition is the Walsh transform [Shanks, 1969].

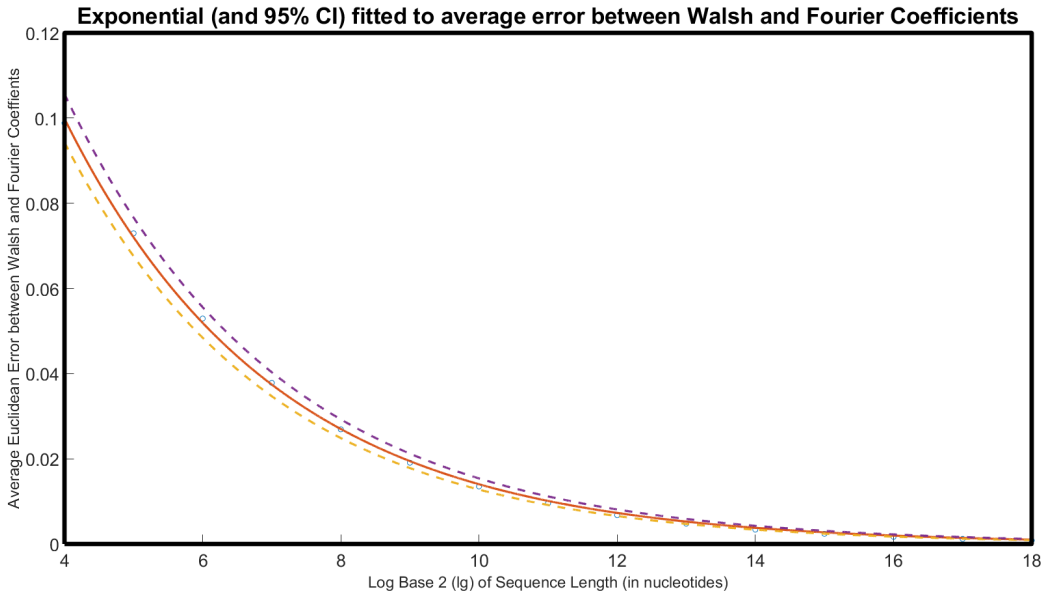


Figure 2.1: Simulation Error Walsh V. Fourier Coefficients for size from 16 to 262,144

Justification for the decision to use a Fourier transform instead of a Walsh transform is validated in a small simulation study. Both the Fourier and Walsh coefficients were computed for 100,000 random sequences each of sizes from 16 to 262,144 in powers of two, and an exponential function regressed through them. The results of the simulation in 2.1 show a swift decay to zero as the sequence size increases. The exponential that was fitted to the

average error between the Fourier and Walsh coefficients at each of the sequence lengths is $f(l) = 0.3695 \cdot e^{-0.3271 \cdot l}$. The coefficient and exponent estimated have 95% confidence intervals of (0.3575, 0.3815) and (-0.3332, -0.3210) respectively. This estimate can be interpreted to be saying that for every log base 2 increase of 1 (each doubling of the length) the average error between the Walsh and Fourier coefficients for the same signal decreases by a factor of 0.7210, or by approximately 30%. Due to this exponential reduction in error, it is decided that the Fourier coefficients will induce a very similar if not isomorphic distance structure among examined sequences. Therefore, instead of applying a Walsh transform, the Fourier transform is used.

2.1.1. Encoding & Transforming Genomic Sequences

Genomic/proteomic signals when digitized can be stored in ASCII encoded files in FASTA/Q format. FASTA stands for Fast-All [Lipman and Pearson, 1985], in reference to the storage capabilities of the format for any alphabet based sequence. For biomolecular sequences these are presented in enormously dense and ultra resilient storage systems such as DNAs and RNAs. The harvested organic material, having been analyzed by sequencing technology, [Metzker, 2010] [Slatko et al., 2018] yields a NT sequence, stored either directly in FASTA/Q or other formats, such as intensity images for micro-array data.

Sequenced genomic data is processed using bioinformatic procedures toward a particular aim. If assembly or alignment is the researcher's goal, a sequence assembler or aligner may be used. An assembler is a tool that builds a long sequence from fragmented shorter sequences. An aligner uses a reference sequence (sometimes built previously using an assembler) and places reads obtained from a sequencer along the reference. Popular aligners for regular DNA sequencing include: HISAT, [Kim et al., 2015] HISAT2[Kim et al., 2019], bowtie2 [Langmead and Salzberg, 2012], and BWA [Li and Durbin, 2009]. Some special purpose aligners for handling sequencing data from NT-conversion, such as HISAT3N [Zhang et al., 2020] have also been developed.

Alignment is a computationally expensive task. If the research task does not require alignment, such as phylogenetic analysis, then the power spectra (PS) can be generated for the sequence without an alignment. First, an encoding strategy must be determined. This corresponds with a decision to use four-vectors or two-vectors [Voss, 1992b] to assign numerical outcomes to the presence or absence of a specific NT. This decision does not always result in contesting relative distance calculations further down the line. It was suggested that encoding using two-vectors with a strategy picked to attempt to "balance" the signals will provide better PS representations [Yin and Yau, 2015]. The mapping into four-vectors is shown in Equation 2.1, and one possible two-vector encoding is shown in Equation 2.2.

$$\begin{aligned} V_4[\cdot] \equiv \{A, C, G, T/U\} \mapsto \\ \{[1,0,0,0]^T, [0,1,0,0]^T, [0,0,1,0]^T, [0,0,0,1]^T\} \end{aligned} \tag{2.1}$$

$$\begin{aligned} V_{2AT}[\cdot] \equiv \{A, C, G, T/U\} \mapsto \\ \{[0,-1]^T, [-1,0]^T, [1,0]^T, [0,1]^T\} \end{aligned} \tag{2.2}$$

The PS, as described by Singleton in 1969 [Singleton, 1969], is computed for each row of the encoded genomic signals (four times for V_4, and two for V_{2AT}). The resulting PS are then averaged together by frequency position, providing a single PS of the same length as the initial signal. If this PS is to be compared to others in an ensemble of signals, a direct comparison method is the computation of the Euclidean distance between the PS.

During the differentiation of genomic sequences, when examining a set of raw data, a researcher might find that the lengths of the genomic sequences of the sample are different. This means that the typical Euclidean distance procedure is not directly applicable to the PS. Instead, the PS may be stretched from smaller sequence lengths to the length of the longest sequences in order to allow for a comparison. When working with the PS, this can be done by applying several different kinds of procedures, such as an even scaling procedure [Yin et al., 2014]. The procedure scans through the shorter sequence taking either a single

PS coefficient or a pair of coefficients, and averages them to produce a signal of the same length as the longer sequence.

Formally, let A_n denote the original power spectrum of length n, and A_m denote the extended power spectrum of length m, where $m > n$ and $k = 1, 2, \ldots, m$ indexes the scaled signal. The even scaling operation from A_n to A_m is given by

$$A_m(k) = \begin{cases} A_n(Q) & \text{if } Q \in Z^+ \\ A_n(R) + (Q - R)(A_n(R + 1) - A_n(R)) & \text{if } Q \notin Z^+ \end{cases} \tag{2.3}$$

where $Q = \frac{kn}{m}$ and $R = \lfloor \frac{kn}{m} \rfloor$ [Yin and Yau, 2015]. The *even scaling procedure* is extensible to signal length differences of up to one half the size of the smaller signal without providing substantive loss in terms of information content. The procedure may be applied systematically in cases where the length difference exceeds this boundary.

A signal flow diagram describing the information stream from the genetic material to the produced *evenly scaled* PS is represented in Figure 2.2. The *evenly scaled* PS are the numerical summaries of the sequences to which the novel filtering procedures are applied. These three filtering approaches are introduced in the next section, and applied to SARS-CoV-2 viral genomes in that which follows it.

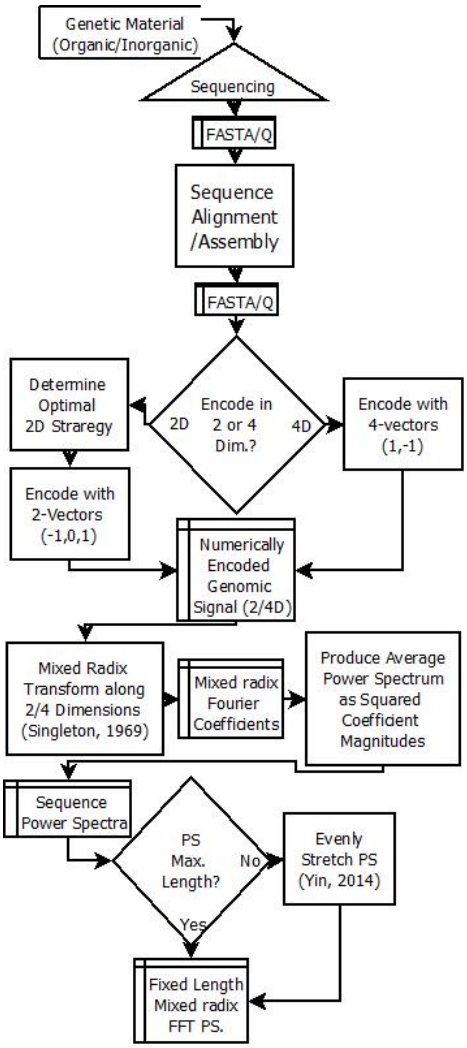


Figure 2.2: Flow Diagram of Genomic Fourier Analysis

2.2. Data Overview

There are approximately 32,000 nucleotides and 29 key protein-encoding genes within the complete SARS-CoV-2 viral genome [Corum and Zimmer, 2020]. Like most viruses, SARS-CoV-2 mutates over time as it comes into contact with various hosts and other pathogens. To date, several variants, such as the Indian variant (B.1.617.2) -δ, the UK variant (B.1.1.7) - α, the South African variant (B.1.351) - β, and the Brazilian variant (P.1) - γ have been identified [Centers for Disease Control and Prevention, 2021].

Data for SARS-CoV-2 sequences is available through the GISAID Initiative [Elbe and Buckland-Merrett, 2017], an organization which maintains records of viral genomes submitted by an international collective of sequencing laboratories. The sequences are maintained in standard multi-FASTA format [Lipman and Pearson, 1985], with information on the submitting lab and capture date included in the sequence headers. All sequences that were submitted to the database had the same general header structure from which the date of submission and location of submission were retrieved by regular expression matching. The initial GISAID multi-FASTA file contains information from 70 submitting laboratory locations, which are coarsened into eight submitting regions according to Table 2.1.

Table 2.1: Regional breakdown by Submitting Locations

Africa	**East Asia**	**Europe**	**Middle East**	**North America**	**Oceania**	**South America**	**West Asia**
Algeria	Beijing	Austria	Turkey	USA	Australia	Brazil	Bangladesh
Egypt	Chongqing	Belgium	Saudi Arabia	Puerto Rico	New Zealand	Chile	Cambodia
South Africa	Fujian	Czech Republic	Kazakhstan	Guam	Indonesia	Colombia	India
DRC	Guangdong	Denmark	Iran	Mexico	Malaysia	Costa Rica	Nepal
Gambia	Hangzhou	Finland	Israel			Uruguay	Sri Lanka
Senegal	Hong Kong	France	Kuwait				Vietnam
	Jiangsu	Georgia					Thailand
	Jiangxi	Germany					
	Jingzhou	Greece					
	Shandong	Spain					
	Shenzhen	Sweden					
	Sichuan	Hungary					
	Taiwan	Portugal					
	Tianmen	Poland					
	Wuhan	Russia					
	Yunnan	Romania					
	Zhejiang	Slovakia					
	Lishui	Italy					
	Japan						
	Guangzhou						

The Fourier coefficients are shown to provide identifiable information for each of 1,397 SCV2 genomes that were considered *complete* and *high-coverage*. Complete means that the sequence length is $\geq$ 29,000 nucleotides and high coverage indicates having a total proportion of non-identifiable nucleotides (displayed as "N", or another single character other than the four "A","C","G",and "T") for less than 1% of the total number of nucleotides in the sequence. Furthermore, only sequences with available patient status and containing 0.05% or fewer unique mutations occurring in the produced amino-acid chains were used. All sequences were submitted on or before the data-capture date of July 18, 2020.

The classification of sequences into the regions in Table 2.1 is assessed by the application of statistical modeling and supervised learning methods on the produced Fourier coefficient power spectrum. Of high interest is the ability of the Fourier coefficients to classify correctly the sequences by geographic origin. The distances provided by a comparison of sequence PS are also compared to distances determined by more traditional methods such as Jukes-Cantor analysis [Jukes et al., 1969] and p-distance [Saitou and Nei, 1987], comparing the correlation of all pairs of distances computed by the various methods.

The Fourier Coefficients for each of the 1,397 sequences were calculated and scaled to the same length for the full set [Yin et al., 2014]. Visualization and clustering techniques were applied to the resulting coefficients. Visualization procedures included TSNE [van der Maaten and Hinton, 2008], and PCA [Hotelling, 1933]. In addition, several different supervised learning techniques were used to classify PS from the sequences into geographic region, the results of which are listed in Table 2.2. The timing footprint in seconds for CPU time for each of the methods is also reported.

Classification using the Euclidean distance between Fourier spectra was compared to distances produced by the Jukes-Cantor and other procedures[Jukes et al., 1969]. The Jukes-Cantor method requires MSA prior to operation, followed by an estimation of the substitution procedures among all sequences to calculate pairwise distances; therefore, it takes much longer to complete than do any of the methods listed in Table 2.2.

Table 2.2: Classification Validation Metrics with Fourier Coefficients (Using 10-fold cross validations)

Classification Scheme	Overall-Accuracy	Average Sensitivity	Average Specificity	CPU Time (s)
ECOC-SVM	0.4624	0.8072	0.2065	2547
Random Forest (25 Trees)	0.7802	0.6019	0.9503	238
Random Forest (50 Trees)	0.7881	0.6195	0.9523	473
Random Forest (100 Trees)	0.7953	0.6210	0.9534	938
Random Forest (500 Trees)	0.7967	0.6238	0.9335	4676
Random Forest (1000 Trees)	0.7996	0.6243	0.9544	9431
Multinomial Logistic Regression (50 Coefficients)	0.3958	0.1277	0.8099	10792
Multinomial Logistic Regression (100 Coefficients)	0.3257	0.1328	0.7583	44942
Neural Network (1 Hidden Layer, 100 Neurons)	0.6707	0.4640	0.9209	2245
Neural Network (1 Hidden Layer, 250 Neurons)	0.6779	0.4982	0.9249	5267
Neural Network (1 Hidden Layer, 500 Neurons)	0.6707	0.4827	0.9217	12255
Neural Network (1 Hidden Layer, 1000 Neurons)	0.6521	0.4722	0.9164	23012
Neural Network (2 Hidden Layers, 250, 150 Neurons)	0.6679	0.4642	0.9201	3905
Pseudo-Quadratic Discriminant Analysis (1000 coefficients)	0.1102	0.2801	0.7495	49
Pseudo-Quadratic Discriminant Analysis (3000 coefficients)	0.073	0.1637	0.7637	231
Pseudo-Quadratic Discriminant Analysis (5000 coefficients)	0.0709	0.1536	0.7640	966

Among the methods listed in Table 2.2 the most accurate procedure overall is random forest [Breiman, 2001]. Furthermore, the table shows that there is only marginal improvement in the method's accuracy as the number of trees in the forest is increased beyond 500 to 1000. The random forest models also showed the highest average sensitivity and specificity from among the models tested. These metrics in addition to the modest training time for the random forest indicated that this classifier is likely to produce the most accurate results, and may be trained in a reasonable amount of time.

2.3. Methods and Procedures

In this section information about the procedures applied to the power spectra (PS) is provided. An introduction to the visualization methods used in this work is followed by a discussion of the transform and encoding strategies. The section concludes by offering suggested filter designs for the PS that allow for further size reduction with retention of vital information. When dealing with a high-dimensional data set, particularly in clustering applications, the information contained in multiple dimensional space is not readily visible to a human observer. We can use graphics programs to assist in drawing three dimensional plots on a two dimensional surface, and there are even some plotting procedures for virtual reality tools which allow for the production of a complete 3D plot [Nagel et al., 2001], but anything

beyond this is not within the human sphere of perception. Therefore it becomes necessary frequently to summarize high dimensional data in which the resultant plots are representable on a 2 Dimensional Surface The visualization of high-dimensional data is a very active area of research, with many different novel and exciting approaches in production. In this section we focus on one such technique known as TSNE, a probabilistic approach to placing multivariate data (or objects) into a visible space where relative distance is displayable, [Hinton and Roweis, 2002].

TSNE may be thought of as a form of dimension reduction insofar as it projects an image of data in high dimensional space onto a low dimensional space while maintaining key information. Other classical approaches to dimension reduction include: PCA whereby selecting the first few components a view of the entire collection of variables associated with a point may be represented and *Multidimension Scaling* (MDS) which seeks to minimize a function called the "strain" representing the relationship of a pairwise distance matrix [Mardia, 1978]. Stochastic Neighbor embedding is a slightly different approach which seeks to use the probability that data points would select one another as neighbors in a lower dimensional space. In SNE, the probability that a data point i would select j as its nearest-neighbor, p_{ij} is calculated by taking into account a distance measure between the two points.

$$p_{ij} = \frac{e^{-d_{ij}^2}}{\sum_{k \neq i} e^{-d_{ik}^2}}$$

A second set of *induced* probabilities, based on the Gaussian distribution originally are described by the following formula:

$$q_{ij} = \frac{e^{-||\boldsymbol{y_i} - \boldsymbol{y_j}||^2}}{\sum_{k \neq i} e^{||\boldsymbol{y_i} - \boldsymbol{y_k}||^2}}$$

In this formula the reduced dimension representation is symbolized as the vectors $\boldsymbol{y_i}$ for each of the i objects. These values are determined by minimizing the gradient of a cost function between the two different probabilities using the Kullback-Liebler divergence.

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \min_{\boldsymbol{y}_{\boldsymbol{i}}} \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The more recent iteration of SNE is known as TSNE [van der Maaten and Hinton, 2008]. The only difference between this and the classical method of SNE is that the embedded distribution is modelled using the t distribution with one degree of freedom (*Cauchy*) rather a Gaussian distribution with variance 0.5. Hence, the only modification of the above formulae has to do with a change in the induced probabilities, which is:

$$q_{ij} = \frac{(1 + ||\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{j}}||)^{-1}}{\sum_{k \neq i} (1 + ||\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{k}}||)^{-1}}$$

This method is coded in R [Donaldson, 2016], as are many packages to display the natural clustering behaviour of the high dimensional data ($\approx$ 32,000 dimensions in this case) in a lower dimensional space (2/3 dimensions) [Konopka, 2020] [De Leeuw and Mair, 2009] [Kassambara and Mundt, 2020].

The Fourier coefficients may be considered as a series of averages of different resolution samples of the same genetic signals with progressively lower sampling rates, decreasing from a_0,which represents full resolution of the signal, to frequencies that sample every other base, every third base, fourth base, etc. Reconsider the encoding of a genomic sequence, and suppose that a sequence of size N where $i = 1, \ldots, N$ is being encoded. l_i corresponds to the nucleotide at location i. Then consider the sequence A_i, which is defined by an indicator function that compares the nucleotide at base i to the nucleotide Adenine. This may be expressed as $A_i = \mathbb{1}(l_i = \text{`A'})$. Where,

$$A_i = \begin{cases} 1, & \text{If } l_i \text{ is an Adenine.} \\ 0, & \text{If } l_i \text{ is not Adenine.} \end{cases}$$

Thus, the resulting sequence of A_i is a simple discrete-index signal that is representative of the presence of Adenine along the genetic signal of interest down to the limit of resolution. The nature of this classification of the data into either present or absent results in a discrete

index signal, which is truly discrete with jumps occurring between zero and one.

Consider further the multivariate time series B_i, again for $i = 1, \ldots, N$ where B_i is the result of a mapping of each singular point l_i onto a binary vector of length four $B_i = b(l_i)$ where:

$$b : \{\text{A, C, G, T}\} \mapsto \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\} \equiv \{\text{A, C, G, T}\} \mapsto \{0, 1, 2, 3\}$$

That this representation is not compressed, and it will be shown later that b may be expressed as a different mapping to achieve a sparser encoding that still retains distinct nucleotide information. For now, however, return to the above encoding function b and the resulting series of B_i 4-vectors which it produces over l_i. Then $\boldsymbol{B} = [\boldsymbol{B_1}, \boldsymbol{B_2}, \ldots, \boldsymbol{B_N}]$ where the $(k, i)^{\text{th}}$ element is denoted $b_{k,i}$, for $k \in \{1, 2, 3, 4\}$. A series of summary statistics for each row of the matrix may be taken to show the composition of each particular nucleotide over the entire sequence, and subsets of the sequence representing lower frequencies. The Fourier transform expressed in section 2 is the continuous domain form. The discrete version of the Fourier transform, called the Discrete Fourier Transform (DFT) replaces this integral with a sum so that the transform becomes:

$$a_n = \frac{1}{N} \sum_{i=0}^{N-1} x(i) e^{-jn\omega_0 i} = \frac{1}{N} \sum_{i=0}^{N-1} x(i) e^{-jn(\frac{2\pi}{N})i} \quad \text{If n} > 0$$

$$a_0 = \frac{1}{N} \sum_{i=0}^{N} x(i)$$

In the above, the sequence of values $\{a_n\}$ for $n = 0, \ldots, N-1$ represents the coefficients of the DFT. The function $x(\cdot)$ is a discretely indexed function that may either have a known analytical form, or simply be a sampled signal, which is sampled at regular and discrete time points. Here, the symbol j is used to denote the imaginary number ($j^2 = -1$). n, the

subscript of the desired coefficient as well as a multiplier of the frequency, is the value which generates sines of higher and higher frequency.

In examining the form of the Fourier transform, it can be seen that the first coefficient is an overall average of the information present in each signal. The relationship of the average of the nucleotides corresponds to the probability of randomly observing a nucleotide in the sequence. Below is a theorem stating this, following the theorem is a short proof.

Theorem 2.3.1 *The first Fourier coefficient of a binary sequence generated by an indicator function for a particular nucleotide on a genomic sequence is equivalent to the probability of selecting a location completely at random on the genome where that same nucleotide is present.*

Proof Consider selecting a position Z completely at random. This means that the position is sampled according to a discrete uniform distribution. It becomes clear that the probability distribution may be expressed:

$$P(Z = i) = \begin{cases} \frac{1}{N} & \text{for } i \in \{x \in \mathbb{Z}^+ | 0 \leq x \leq N-1\} \\ 0 & \text{for } i \notin \{x \in \mathbb{Z}^+ | 0 \leq x \leq N-1\} \end{cases}$$

Where here $\mathbb{Z}^+$ refers to the non-negative whole integers. Now suppose we are interested in the random variable A_Z. This is a random variable as it is indexed by a random variable, the valuation of this random variable is A_i, which is either 0, or 1 depending on whether there is an adenine present at location i. Therefore $\mathcal{S}(A_Z) = \{1, 0\})$ and we are interested in $P(A_Z = 1) = \pi_A$. As we are equally likely to select any index, i, along the entire sequence, we may say that $P(A_Z = A_i) = \pi_A^{A_i} \cdot (1 - \pi_A)^{1-A_i}$. For a given genetic sequence we observe all A_i for $i \in \{1, \ldots, N\}$ Therefore at each location $i, P(A_i = 1) = A_i$. Then

$$P(A_Z = 1) = P(Z = i \cap A_i = 1) = P\{(Z = 1 \cap A_1 = 1) \cup (Z = 2 \cap A_2 = 1) \cup \cdots \cup (Z = N \cap A_N = 1)\}$$

It is the case for

$$\{(i, j) \in \{1, \ldots, N\} \times \{1, \ldots, N\} | i \neq j\} \text{ that } (Z = i \cap Z = j) = \emptyset$$

$$\implies \{(Z = i \cap A_i = 1) \cap (Z = j \cap A_j = 1)\} = \emptyset$$

$$\begin{aligned}\pi_A &= P(A_Z = 1) \\ &= P(Z = 1) \cdot P(A_1 = 1) + P(Z = 2) \cdot P(A_2 = 1) + \cdots + P(Z = N) \cdot P(A_N = 1)\end{aligned}$$

$$= \sum_{i=1}^{N} P(Z = i) \cdot P(A_i = 1) = \sum_{i=1}^{N} \frac{1}{N} \cdot A_i = \frac{1}{N} \sum_{i=1}^{N} A_i.$$

Indexing instead by i, and letting $x(i) = \mathbb{1}(l_i = \text{'A'})$ we have:

$$\pi_A = \frac{1}{N} \sum_{z=1}^{N} A_z = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(l_i = \text{'A'}) = \frac{1}{N} \sum_{i=1}^{N} x(i) = a_0. \quad \blacksquare$$

The encoding strategy for genetic sequences that is utilized in this research is related to the previously discussed radix-4 encoding, and was initially described in [Yin and Yau, 2015]. The strategy represents the four-category genomic sequence using three valued two-vectors. That is, instead of using two-valued four-vectors which oscillate between the values one and zero, two three-valued vectors which contain only the values -1, 0 and 1 are used. This approach, known as 2D Voss encoding [Voss, 1992a, Yin and Yau, 2015], improves performance of the DNA similarity analysis method is to create a 2D representation of the sequence instead of a 4D representation. It is the 2D representation, described ib Equation 2.4 that we

employ in this research.

Suppose the mapping β defined as

$$\beta(A) = [0, -1]', \quad \beta(T) = [-1, 0]', \quad \beta(C) = [1, 0]', \quad \beta(G) = [0, 1]' \tag{2.4}$$

Thus, the DNA sequence is defined as a 2D matrix ν as $\nu(n) = [\eta_1(n), \eta_2(n)]' = \beta(\alpha_n)$, where $\alpha \in \{A, C, G, T\}, n = 0, 1, 2, \ldots, N-1$, for each sequence of length N. The 2D approach allows for eight possible encodings, which produce three unique power spectra. Of the eight, only two produce signals balanced in the quantity of zero vs. non-zero content. Equation 2.4 displays the AG encoding. The encoding which produces the most balanced (and hence stationary) signal is driven by data-specific ratios [Yin and Yau, 2015]. Both 2D and 4D procedures are provided as a MatLab implementation in [Yin, 2020], and an R Implementation [Thornton, 2021].

To compute pairwise distances (Euclidean or otherwise) among the power spectra, they must first be scaled to be of the same length. Some have suggested the use of partial spectra from a few beginning frequencies [Wu et al., 2000], [Wang et al., 2013],
[Rafiei and Mendelzon, 1998]; but this leads to a loss of information in the comparison. A different way of managing the lengths of the DFT power spectra is to alternate selection of one or the average of two consecutive elements in the shorter signal, which has the effect of "stretching" the shorter signal to the same length as the longer signal. Once Equation 2.3 is applied to the power spectra of the genomic signals, the Euclidean distance between each pair of sequences is calculated.

2.3.1. Comparison to Other Distance Metrics

In this work several methods for determining the distances between genetic sequences accounting for gaps in the sequences built into Matlab are used. The first method, known as *p-distance*, is a measure from zero to one which represents the proportion of locations at which the two sequences being compared differ from each other. The p-distance is a

procedure that requires two sequences being compared to be aligned with each other so that the nucleotides at each position (loci along the first sequence) may be compared with those at the corresponding loci in the second sequence [Paradis and Schliep, 2019]. The second, the Jukes-Cantor method, is a simple logarithmic function of p-distance. The Jukes-Cantor distance is given by $-\frac{3}{4} \cdot \log((1-p) \cdot \frac{4}{3})$, where p is the p-distance.

No obvious relationship between the Fourier coefficient distance and the Jukes-Cantor/p-distance methods exists. Any relationship discovered between these two methods would indicate that both contain similar information. To examine this relationship, Pearson's correlation coefficient is calculated between the three sets of distances: Fourier coefficients vs. Jukes-Cantor, p-distance vs. Jukes-Cantor, and Fourier coefficients vs. p-distance. A null hypothesis significance test for $H_0 : \rho = 0$ versus $H_A : \rho \neq 0$ indicates evidence for an association among the three correlation measures. The p-value for the correlation test between the p-distance and the Fourier Coefficient Distance is 0.0197, and that between the Jukes-Cantor distance and the Fourier coefficient distance was very small and displayed as 0. This result indicates that there is some shared information among the three distances.

The distances discussed in this section so far have been ones which require multiple sequence alignment (MSA). There are many MSA algorithms. In this work, Clustal W software was used to perform MSA on the complete set of 1,397 sequences in this study to allow for subsequent analysis using the classical substitution models
[Larkin et al., 2007].

In contrast to these post-MSA procedures, one frequently used alignment-free method for characterizing data is known as the k-mer distance calculation. There are several variations of this procedure but these all typically begin with counting the k-mers in a file of a given size [Röhling et al., 2020] [Allman et al., 2017] [Benoit et al., 2016] [Wen et al., 2014]. There have been a few software packages developed for rapidly computing these frequency or count vectors. One popular program known as Jellyfish
[Marcais and Kingsford, 2012] implements a k-mer counting routine that allows for parallel processing, making it a very rapid and accurate counter. In this research however the case

study genomes were small enough that a basic k-mer counting sub-routine is implemented in the R-package developed for this study [Thornton, 2021]. In these distance calculations, the K-mers, where $k \in \{1, 2, 3, 4, 5\}$ were counted for each sequence, and the resultant frequency vector is compared to other sequences via the Euclidean distances between the frequency vectors. A complete listing of the phylogenetic distance procedures used in this work is provided in A.1.

2.3.2. Filtering Procedures

Since the initial PS may be large, subsets of PS coefficients produced by a few techniques, with special focus on preserving relative distances captured by the entire unfiltered genomic PS might be desired by a researcher. To this end, non-parametric data-driven filters are applied to a subset of SARS-CoV-2 virus genomes.

These approaches to filtering the PS were designed to determine appropriate subsets which might be used to numerically summarize genomic signals such as DNA and RNA while retaining the distance information provided by the fully unfiltered PS. In this way, they reduce the extent of the PS analyzed while attempting to maintain relative orderings of pair-wise distances among the sequences analyzed.

Three approaches, *Minimal Variance Filtering* (MVF), *Automated Filter Learning* (AFL), and *Maximal Variance Principal Components Filters* (MVPCF), are suggested and their utilities examined. Retention of pair-wise distance orderings is assessed with Spearman's Rank Correlation coefficient (ϱ) to measure the linear relationship between Euclidean distances of ranks among full unfiltered PS and reduced filtered PS.

As the techniques described here are non-parametric and data-driven, either a representative sample for construction of the filters must be selected among those available, or the entire set may be used to derive an ensemble-specific set of filters. This indicates that when new genomic signals are collected, the same filters may be applied as are or reconstructed with the new signals. Filters built with new signals may not be identical to those produced

by the originals. Thus, the filters may be updated whenever a new signal is added to the ensemble.

With the specification of the percentage of PS coefficients to filter, the MVF technique can exclude those PS coefficients which vary the least amongst the sample. The variance of each of the $j = 1, 2, \ldots, m$ elements of the *evenly-scaled* PS ensemble for either the entire sample, or some representative subsample is computed. The supplied q, the fraction of the coefficients by which to reduce the original PS, is used to determine which variance should be the cutoff in the data, in accordance with Equation 2.5.

$$
\begin{aligned}
v_j =& \frac{\sum_{i=1}^{N} \left(S_{ij} - \frac{\sum_{i=1}^{N} (S_{ij})}{N} \right)^2}{N-1} \\
\eta^* =& \left\{ v_j \left| \sum_{k=1}^{m} \mathbb{1} \left(v_j \leq v_k \right) = \lfloor m \cdot q \rfloor \right. \right\} \\
S_{il}^* =& \left\{ S_{ij} \middle| v_j \geq \eta^* \right\}
\end{aligned}
\tag{2.5}
$$

The variance filtering approach is applied to genomic PS coefficients S_{ij}, for coefficient $j = 1, 2, \ldots m$ of sample $i = 1, 2, \ldots, N$, to produce filtered PS S_{il} for $l = 1, 2, \ldots, \lfloor m \cdot q \rfloor$. The filtered PS S_{il} are produced by selecting only those PS coefficients from the unfiltered PS that have the largest across-sample variances.

This filter design may be considered as semi-analogous to that of a *matched filter* from traditional signals processing. As a matched filter seeks to exploit a template, so too does the MVF technique. Once a threshold determination (η^*) is made from the composite variances of the coefficients each PS is matched to it by selection only if that coefficient has a higher variance. The MVF design does not inspect the composite variances at lagged PS alignments, and hence is only similar to the lag-zero *matched filter* [Woodward, 2014] [Turin, 1960] [North, 1963] [Jaynes, 2003]. AFL is a less hands on approach than MVF, and provides a set of unique linear combinations of the entire genomic PS instead of a binary filter. These filters are learned automatically by a 1-D Convolutional Neural Network

(CNN) applied to the full PS [Zhang et al., 1988]. The CNN is designed to classify important attributes of the original genomic sequences. The characteristics classified by the CNN may vary depending on application. For example, in the SARS-CoV-2 case study, the labels were regional submission data that was extracted from the headers provided by GISAID. A schematic of the Machine Learning (ML) approach is shown in Figure 2.3 and depicts the CNN layer learning filter coefficients for inputted PS.

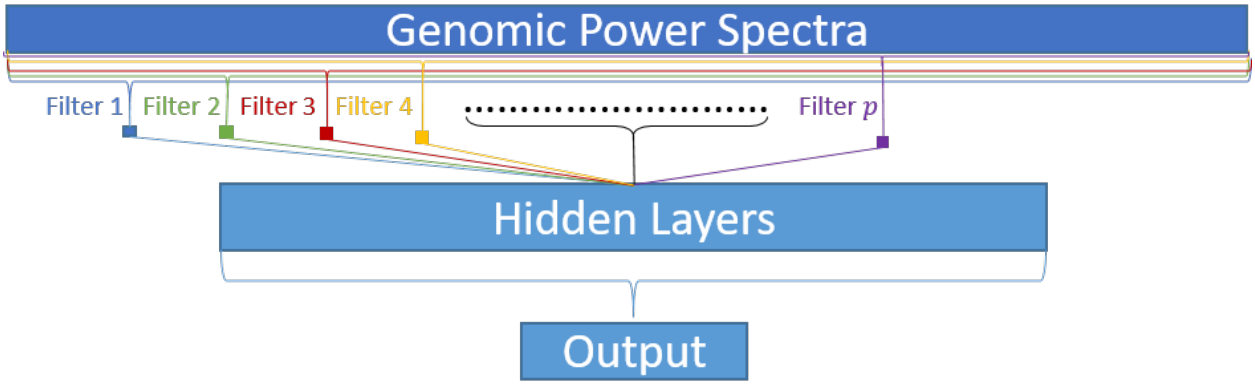


Figure 2.3: Schematic of Machine Learning approach for automatic filter learning

As with the previous approach, the size of the filtered PS can be selected by the researcher. In contrast to MVF, the resultant set of filtered PS coefficients actually consist of linear combinations of the unfiltered PS coefficients. In addition, AFL is dependent on having supervised labels for the sequences used to construct the filters. These labels will determine the kinds of filters constructed and the information that they emphasize; hence, aspects of the original PS that help differentiate the elements of the sample on the chosen labels will be extracted by the filters. The convolutional network layer is applied to the entire length of the PS, p, which is the number of filters to be learned by the CNN.

The constructed filters also depend on the architecture of the hidden layers between the output and the convolutional layer. In the results section we will look at a specific architecture that learns filters by attempting to classify the sequences by region of submission. A notable drawback of this approach is its dependence on a known label that can either be extracted from the data or supplied by a user.

MVPCF provides filters in the same style as those automatically learned in AFL with the advantage of stronger correlation to distances calculated from the unfiltered PS. For the largest subset of PS coefficients given N samples, the principal components of the N highest variance PS coefficients in the sample are taken as the filtered PS. When the number of samples, N, from which the filter is constructed is less than the length of the PS, m, this technique selects the N highest variance PS coefficients, otherwise the entire PS is utilized. In this case, the selection of the length of the filtered PS is made by choosing only the first k principal components.

Let $\boldsymbol{X}$ be the matrix containing PS of all samples from which we want to construct a filter. Each row $i = 1, 2, \ldots, N$ represents a signal, and each column $j = 1, 2 \ldots, m$ represents the j^{th} PS coefficient. A pre-filtering matrix $\boldsymbol{H}$ may be determined from η^* in Equation 2.5, such that $\boldsymbol{H}$ is of size $m \times k^*$, where k^* is at most N. Each column of $\boldsymbol{H}$ selects the single PS coefficient from $\boldsymbol{X}$ of variance greater than η^* and less than the previously selected PS coefficient, such that $\boldsymbol{XH} = \boldsymbol{X^*}$ is of size $N \times k^*$; thus, $\boldsymbol{H}$ contains only the maximal variance coefficients of $\boldsymbol{X}$. The singular value decomposition (SVD) of the matrix $\boldsymbol{X^*}$ may be written $\boldsymbol{X^*} = \boldsymbol{U\Sigma W}^T$ with score matrix $\boldsymbol{T} = \boldsymbol{X^*W} = \boldsymbol{U\Sigma}$. Then a post filtering matrix $\boldsymbol{K}$ may be applied to the score matrix, such that $\boldsymbol{K}$ is of size $k^* \times k$ and therefore $\boldsymbol{TK} = \boldsymbol{X^*WK} = \boldsymbol{U\Sigma K} = \boldsymbol{F}$ is the $N \times k$ *maximal variance principal components filtered* representation of the initial PS matrix $\boldsymbol{X}$.

2.3.3. Complex Gaussian Model for Fourier Coefficients

Now a parametric hypothesis test that uses the complex normal distribution is introduced. The test also relies on the assumption of second-order stationarity to directly compare samples of signal autocorrelations. The correlation structures are compared by inspecting summary statistics computed from their groupings.

The asymptotic distribution of the Fourier coefficients of a signal, as the signal length tends to infinity, is known as the multivariate complex normal distribution [Brillinger, 2001]. That is, the distribution of each of the coefficients is marginally a complex normal distribution as shown in 2.6.

$$\boldsymbol{d}_X^{(T)}(\lambda) = \sum_{t=0}^{T-1} X(t)e^{-i\lambda t} \implies \boldsymbol{d}_X^{(T)}(\lambda_j(T)) \sim N_r^{\mathbb{C}}(\mathbf{0}, 2\pi T \boldsymbol{f_{XX}}(\lambda_j(T))) \tag{2.6}$$

In Equation 2.6 the complex multivariate normal distribution on the complex r-vector Fourier Coefficients of frequency $(\lambda_j(T))$, denoted as $\boldsymbol{d}_X^{(T)}(\lambda_j(T))$,is symbolically represented $N_r^{\mathbb{C}}$. In the case of genetic sequences, the base nucleotide (NT) sequence is first converted into a numerical signal using one of two encoding strategies, which produce either a four, or two vector valued numeric signal ($r \in \{2, 4\}$). In Equation 2.6, $\lambda_j(T)$ is defined in terms of the integers $s_j(T)$, where $\{s_j(T) \in \mathbb{Z} | 0 \leq s_j(T) \leq T\}$ such that:

$$\lambda_j(T) = \frac{2\pi s_j(T)}{T} \quad \lim_{T\to\infty} \lambda_j(T) = \lambda_j \quad \lambda_j = \frac{2\pi s_j}{T} \tag{2.7}$$

$\boldsymbol{f_{XX}}$, the $r \times r$ *spectral density matrix* of X(t) for a given frequency, λ is computed by examination of the $r \times r$ *autocovariance function* $\boldsymbol{c_{XX}(u)}$:

$$\begin{aligned} \boldsymbol{f_{XX}}(\lambda) &= \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} \boldsymbol{c_{XX}}(u)e^{-i\lambda u} \quad -\infty < \lambda < \infty \\ \boldsymbol{c_{XX}}(u) &= \text{cov}(X(t+u), X(t)) \end{aligned} \tag{2.8}$$

where $\boldsymbol{c_{XX}}(u)$ is the *autocovariance function*, a $r \times r$ matrix valued function. Under the

condition of second-order stationarity, which is a key underlying assumption, the autocovariance function is not a function of time. The multivariate complex normal distribution on r-vectors has a distribution that can be expressed in terms of the multivariate normal distribution on $2r$-vectors. Given an r-vector random variable $\boldsymbol{Z} = \boldsymbol{Z}_{\Re} + j \cdot \boldsymbol{Z}_{\Im}$ with sample space $\mathcal{S}_{\boldsymbol{Z}} \equiv \mathbb{C}^r$, the following two distributions, with vector (mean) and matrix-parameters (variance-covariance matrix) $\boldsymbol{\mu_X} = \boldsymbol{\mu}_{\boldsymbol{X}_{\Re}} + j \cdot \boldsymbol{\mu}_{\boldsymbol{X}_{\Im}}$ and $\boldsymbol{\Sigma_{XX}} = \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}_{\Re}} + j \cdot \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}_{\Im}}$ on $\boldsymbol{Z}$ are equivalent, in the case of a r-vector signal X with transform coefficient Z.

$$\boldsymbol{Z} \sim N_r^{\mathbb{C}}(\boldsymbol{\mu_X}, \boldsymbol{\Sigma_{XX}}) \iff \begin{pmatrix} \boldsymbol{Z}_{\Re} \\ \boldsymbol{Z}_{\Im} \end{pmatrix} \sim N_{2r}\left(\begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{X}_{\Re}} \\ \boldsymbol{\mu}_{\boldsymbol{X}_{\Im}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}_{\Re}} & -\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}_{\Im}} \\ \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}_{\Im}} & \boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{X}_{\Re}} \end{pmatrix} \right) \tag{2.9}$$

Recall that the probability density function for a multivariate normal distribution $N_{2r}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is expressed as:

$$\boldsymbol{Q} \sim N_{2r}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff dP(\boldsymbol{Q} \leq \boldsymbol{q}) = \left((2\pi)^{2r}|\boldsymbol{\Sigma}|\right)^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{q}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{q}-\boldsymbol{\mu})} d\boldsymbol{q} \tag{2.10}$$

This formulation of the distribution for second-order stationary signals is used in the subsequent derivation of the statistical hypothesis test based on the Likelihood Ratio Test for observed variance-covariance matrices of the multivariate Fourier coefficients.

Using Brillinger's notation for the Fourier Transform distribution it becomes clear that $\boldsymbol{V_r(\omega)}$ from Equation 1.1 is equivalent to $\boldsymbol{d}_X^{(T)}(\lambda)$ in Equation 2.6. If the assumption of second-order stationarity holds, it can be shown that the asymptotic distribution of $\boldsymbol{d}_X^{(T)}(\lambda)$ converges to the complex normal displayed in Equation 2.6. This is almost certainly not the case for genomic sequences, as the auto-covariance function is likely not time-independent (meaning that the relationship between each point t in the signal and the point u elements ahead or behind does not depend on the point t). That said, this simplifying assumption is critical to *feasibly* make use of the estimated variance-covariance matrix in the test hereafter described. In future works, we will investigate transforms of the initial sequences that may allow this assumption to be more readily met, in addition to performing an analysis of the robustness of the statistical procedure to this assumption.

Suppose that a sample of N values from a r-variate complex normal random variable is taken such that $\boldsymbol{Z_1}, \boldsymbol{Z_2}, \ldots, \boldsymbol{Z_N}$ represent the sample random variables, and $\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_N}$ the observed valuations. Suppose further that it is the case that $\boldsymbol{Z_n} \overset{\text{iid}}{\sim} N_r^{\mathbb{C}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, as is indicated in Equation 2.9. This distribution may also be expressed in terms of the multivariate normal distribution, with mean-vector and variance-covariance matrix expanded as displayed. Let $\boldsymbol{\Sigma_\Re}, \boldsymbol{\mu_\Re}, \boldsymbol{Z_{n\Re}}$ and $\boldsymbol{z_{n\Re}}$ be the real parts of $\boldsymbol{\Sigma}, \boldsymbol{\mu}, \boldsymbol{Z_n}$ and $\boldsymbol{z_n}$ and $\boldsymbol{\Sigma_\Im}, \boldsymbol{\mu_\Im}, \boldsymbol{Z_{n\Im}}$ and $\boldsymbol{z_{n\Im}}$ their corresponding imaginary portions respectively. Furthermore, let the scalars $z_{n\Re l}$ and $z_{n\Im l}$ represent the l^{th} components of the real and imaginary vectors $\boldsymbol{z_{n\Re}}$ and $\boldsymbol{z_{n\Im}}$, then the frequency function of the joint distribution of the sample may be expressed as the argument of the partial differential equation:

$$\begin{aligned}
&\partial P(\boldsymbol{Z_1} \leq \boldsymbol{z_1} \cap \boldsymbol{Z_2} \leq \boldsymbol{z_2} \cap \cdots \cap \boldsymbol{Z_N} \leq \boldsymbol{z_N}) \\
&= \prod_{n=1}^{N} \left((2\pi)^{2r} \left| \begin{pmatrix} \boldsymbol{\Sigma_\Re} & -\boldsymbol{\Sigma_\Im} \\ \boldsymbol{\Sigma_\Im} & \boldsymbol{\Sigma_\Re} \end{pmatrix} \right| \right)^{-\frac{1}{2}} \\
&\cdot \exp\left(-\frac{1}{2} \left(\begin{pmatrix} \boldsymbol{z_{n\Re}} \\ \boldsymbol{z_{n\Im}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu_\Re} \\ \boldsymbol{\mu_\Im} \end{pmatrix} \right)^T \begin{pmatrix} \boldsymbol{\Sigma_\Re} & -\boldsymbol{\Sigma_\Im} \\ \boldsymbol{\Sigma_\Im} & \boldsymbol{\Sigma_\Re} \end{pmatrix}^{-1} \left(\begin{pmatrix} \boldsymbol{z_{n\Re}} \\ \boldsymbol{z_{n\Im}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu_\Re} \\ \boldsymbol{\mu_\Im} \end{pmatrix} \right) \right) \partial \boldsymbol{z_n}
\end{aligned} \tag{2.11}$$

The Maximum Likelihood Estimators of the parameters of the multivariate normal distribution are the vector average of observations and the sample variance-covariance matrix of the sample, which may be computed according to Equations 2.12, and 2.13. Recall that each of the vectors are of length r, and let the final subscripted index represented the l^{th} element of the vector.

$$\widehat{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\boldsymbol{\mu_\Re}} \\ \widehat{\boldsymbol{\mu_\Im}} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \widehat{\mu_{\Re 1}} & \widehat{\mu_{\Re 2}} & \ldots & \widehat{\mu_{\Re r}} \end{pmatrix}^T \\ \begin{pmatrix} \widehat{\mu_{\Im 1}} & \widehat{\mu_{\Im 2}} & \ldots & \widehat{\mu_{\Im r}} \end{pmatrix}^T \end{pmatrix} = \begin{pmatrix} \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{z_{n\Re}} \\ \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{z_{n\Im}} \end{pmatrix} \tag{2.12}$$

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma_\Re}} & -\widehat{\boldsymbol{\Sigma_\Im}} \\ \widehat{\boldsymbol{\Sigma_\Im}} & \widehat{\boldsymbol{\Sigma_\Re}} \end{pmatrix}$$

$$\widehat{\boldsymbol{\Sigma}_{\Re}} = \begin{pmatrix} \widehat{\sigma^2_{\Re 1}} & \widehat{\rho_{\Re 1,2}} & \cdots & \widehat{\rho_{\Re 1,r}} \\ \widehat{\rho_{\Re 2,1}} & \widehat{\sigma^2_{\Re 2}} & \cdots & \widehat{\rho_{\Re 2,r}} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\rho_{\Re r,1}} & \widehat{\rho_{\Re r,2}} & \cdots & \widehat{\sigma^2_{\Re r}} \end{pmatrix} \qquad \begin{aligned} \widehat{\sigma^2_{\Re l}} &= \tfrac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Re l} - \widehat{\mu_{\Re l}}\right)^2 \\ \widehat{\rho_{\Re l,m}} &= \tfrac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Re l} - \widehat{\mu_{\Re l}}\right)\left(z_{n\Re m} - \widehat{\mu_{\Re m}}\right) \end{aligned}$$

$$\widehat{\boldsymbol{\Sigma}_{\Im}} = \begin{pmatrix} \widehat{\sigma^2_{\Im 1}} & \widehat{\rho_{\Im 1,2}} & \cdots & \widehat{\rho_{\Im 1,r}} \\ \widehat{\rho_{\Im 2,1}} & \widehat{\sigma^2_{\Im 2}} & \cdots & \widehat{\rho_{\Im 2,r}} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\rho_{\Im r,1}} & \widehat{\rho_{\Im r,2}} & \cdots & \widehat{\sigma^2_{\Im r}} \end{pmatrix} \qquad \begin{aligned} \widehat{\sigma^2_{\Im l}} &= \tfrac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Im l} - \widehat{\mu_{\Im l}}\right)^2 \\ \widehat{\rho_{\Im l,m}} &= \tfrac{1}{N-1}\sum_{n=1}^{N}\left(z_{n\Im l} - \widehat{\mu_{\Im l}}\right)\left(z_{n\Im m} - \widehat{\mu_{\Im m}}\right) \end{aligned} \tag{2.13}$$

Given a set of K total genomic sequence samples (that is, K groups of genomic sequences), there may be a desire to determine whether the variance-covariance structure of the samples of sequences differ. Taking the Fourier transform of all sequences at a given frequency will allow the maximum likelihood estimation of the variance covariance structure as shown in Equations 2.12, and 2.13, for all of the data. Hence the likelihood function for an MLE over all of the data can be expressed using Equation 2.11 as Equation 2.14.

$$\begin{aligned} \log L\left(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}} \;\; ; \;\; (\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_N})^T\right) &= \ell(\widehat{\boldsymbol{\Sigma}}) \\ &= \frac{N}{4} \cdot \log\left((2\pi)^{2r}|\widehat{\boldsymbol{\Sigma}}|\right) \sum_{n=1}^{N} \begin{pmatrix} \boldsymbol{z_{n\Re}} \\ \boldsymbol{z_{n\Im}} \end{pmatrix}^T \widehat{\boldsymbol{\Sigma}}^{-1} \begin{pmatrix} \boldsymbol{z_{n\Re}} \\ \boldsymbol{z_{n\Im}} \end{pmatrix} \end{aligned} \tag{2.14}$$

Suppose that each of the K groups of genomic signals has a sufficient number of observations to allow for the individual estimation of parameters, and that the number of observations in the k^{th} group is denoted by N_k for $k = 1, 2, \ldots, K$ such that $\sum_{k=1}^{K} N_k = N$, and letting the observed Fourier coefficients vector for a specified element of subgroup k be specified as $\boldsymbol{z_{n_k\Re(k)}}$ and $\boldsymbol{z_{n_k\Im(k)}}$ where n_k references the particular element in the k^{th} subgroup for the real and imaginary parts of the Fourier coefficient respectively. Suppose also that the maximum likelihood estimates of the variance-covariance matrices for each of the K subgroups of data are estimated as $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}$ for $k \in \{1, 2, \ldots, K\}$ with submatrices $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k\Re}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k\Im}}$. Then a second likelihood function which considers each of the subgrouped data points separately may be

constructed using each of the grouped data as shown in Equation 2.15.

$$\ell(\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}},\ldots,\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{K}}) = \sum_{k=1}^{K}\left(\frac{N_k}{4}\cdot\log\left((2\pi)^{2r}|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}|\right)\sum_{n_k=1}^{N_k}\left(\begin{pmatrix}\boldsymbol{z}_{\boldsymbol{n_k}\Re(\boldsymbol{k})}\\ \boldsymbol{z}_{\boldsymbol{n_k}\Im(\boldsymbol{k})}\end{pmatrix}^{T}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}{}^{-1}\begin{pmatrix}\boldsymbol{z}_{\boldsymbol{n_k}\Re(\boldsymbol{k})}\\ \boldsymbol{z}_{\boldsymbol{n_k}\Im(\boldsymbol{k})}\end{pmatrix}\right)\right) \tag{2.15}$$

Note that under the hypothesis (H_0) that $\widehat{\boldsymbol{\Sigma}_{\mathbf{1}}} = \widehat{\boldsymbol{\Sigma}_{\mathbf{2}}} = \cdots = \widehat{\boldsymbol{\Sigma}_{\boldsymbol{K}}}$, Equation 2.15 becomes Equation 2.14. Therefore, we may test this hypothesis with the standard Likelihood Ratio Test (LRT) statistic given in Equation 2.16.

$$\Lambda = -2\cdot\left(\ell(\widehat{\boldsymbol{\Sigma}}) - \ell(\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}},\ldots,\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{K}})\right) \implies \Lambda \sim \chi^2_{K-1} \tag{2.16}$$

This test statistic may then be examined to determine whether or not it falls beyond the researcher's selected critical point into the tail of its χ^2 distribution.

2.4. Results and Analysis

The results of applying these frequency analysis techniques to the SARS-CoV-2 data set described are displayed in this section. First the results of applying clustering procedures for the visualization of the power spectra are provided, in particular, TSNE [Hinton and Roweis, 2002],[van der and canonical variables plots are shown. The visualization procedure is followed directly by a traditional MANOVA procedure [Stevens, 2007]. This is followed by the results of applying supervised learning procedures to both the DFT Power Spectra, as well as K-mer frequency vectors [Wen et al., 2014]. The Correlation between pairwise distances in the data for several classical substitution model approaches such as Jukes-Cantor and Kimura [Kimura, 1980],[Kimura, 1981] as well as those produced with the Fourier PS are examined. The results of filtering the PS and attempting clustering and classification are then discussed.

2.4.1. Clustering and Visualization for SARS-CoV-2 Sequences

To examine whether the scaled spectra for the various SARS-CoV-2 genomes contain information relevant to the differentiation of the geographic location from which they were submitted, a TSNE plot was constructed using the power spectra (Figure 2.4). The points on the TSNE plot are colored by geographic location according to Table 2.1. Note that location on the TSNE plot is fairly well differentiated; however, there are some regions which do not cluster on the scaled Fourier spectra. This might have to do with the spread of the virus to different areas by a particular individual visiting from another geographic region from which an original or a similar and presumably closely related viral sequence is submitted.

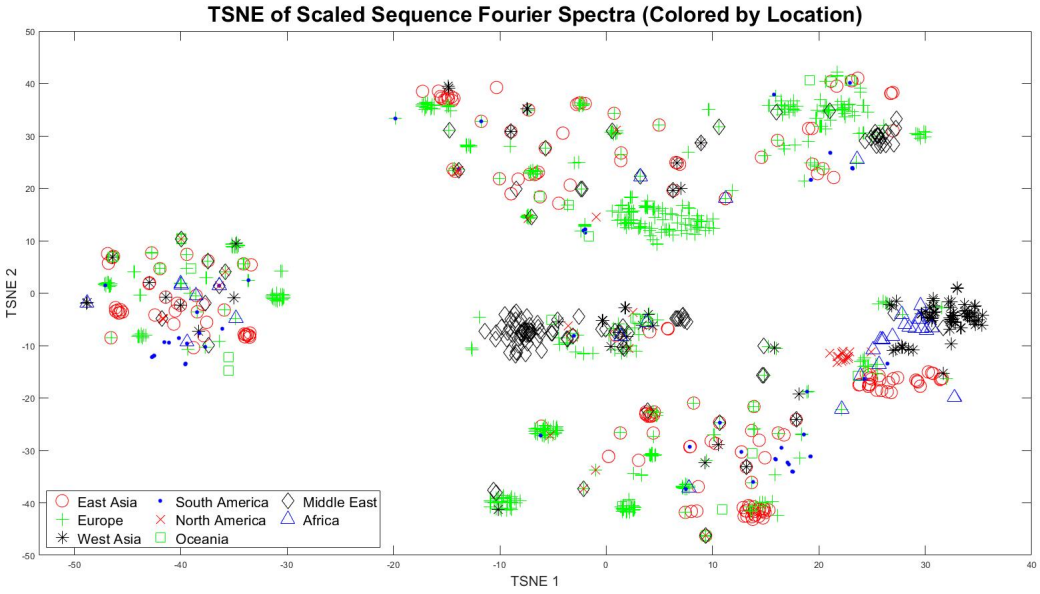


Figure 2.4: Complete Fourier Spectra Visualization: TSNE plot

The tight clusters of particular regional data that are separated by some distance indicate that the actual location of submission of the particular sequence is associated with the viral sequences' harmonic data. A TSNE plot of the first thousand coefficients alone is shown here as well to display that a subset of the coefficients may still be used to illuminate some of the underlying structure of the high-dimensional data (Figure 2.5). We used TSNE plots

to visualize K-mer frequency vectors (1364 dimensions) and Fourier power spectra (33,133 dimensions) euclidean distances in two-dimensional space. The dimension of the Fourier Power Spectra is the same size as the number of coefficients computed and scaled for each of the signals, while that of the K-mer frequency vector is determined by the number of K-mers used, in this case the first 5, (ie. $4^1 + 4^2 + 4^3 + 4^4 + 4^5 = 4 + 16 + 64 + 256 + 1024 = 1364$).

Figure 2.5 shows TSNE plots for two alignment-free (K-mer, and Power Spectra) methods (top row). Each color in the plot represents a different geographic region, as defined in Table 2.1. In the TSNE plot, points representing probabilities p_{ij} that sequence i would have sequence j as its neighbor are colored by geographic location. The p_{ij} are calculated via optimization of the Kullback-Leibler divergence between the distribution of Euclidean distances in high-dimensional space, and then in the low-dimensional space.

TSNE plots for Jukes-Cantor [Jukes et al., 1969] and Kimura [Kimura, 1980] distances are displayed in the bottom row. Both Jukes-Cantor and Kimura distances require multiple sequence alignment (MSA) before distances can be computed. These more classical distances were computed not only as comparisons for clustering and classification, but also to determine whether the Jukes-Cantor, Kimura, K-mer, and FC distances were providing redundant information about the relationships among the genomes.

The results of the visualizations in Figures 2.5 are encouraging, and indicate that both the power spectra and the K-mer frequency vectors are valid numerical summaries for analysis of the geographic origin of a particular virus sample.

The patterns in the upper left TSNE plot (K-mer distance) are indicative of redundancy in the data (*i. e.* given five-mer frequencies, the frequencies of all lower mers can be computed).

One can see that the colors tend to cluster together, particularly for the K-mer and FC distances; however, there is some mixing. The plot constructed from the Fourier spectra (top right) shows distinct clusters of sequences, particularly for West Asia (bright pink) and Europe (olive green). In addition, these two clusters of points are separated from each other. The cluster for West Asia is closest to the cluster for the Middle East, which makes

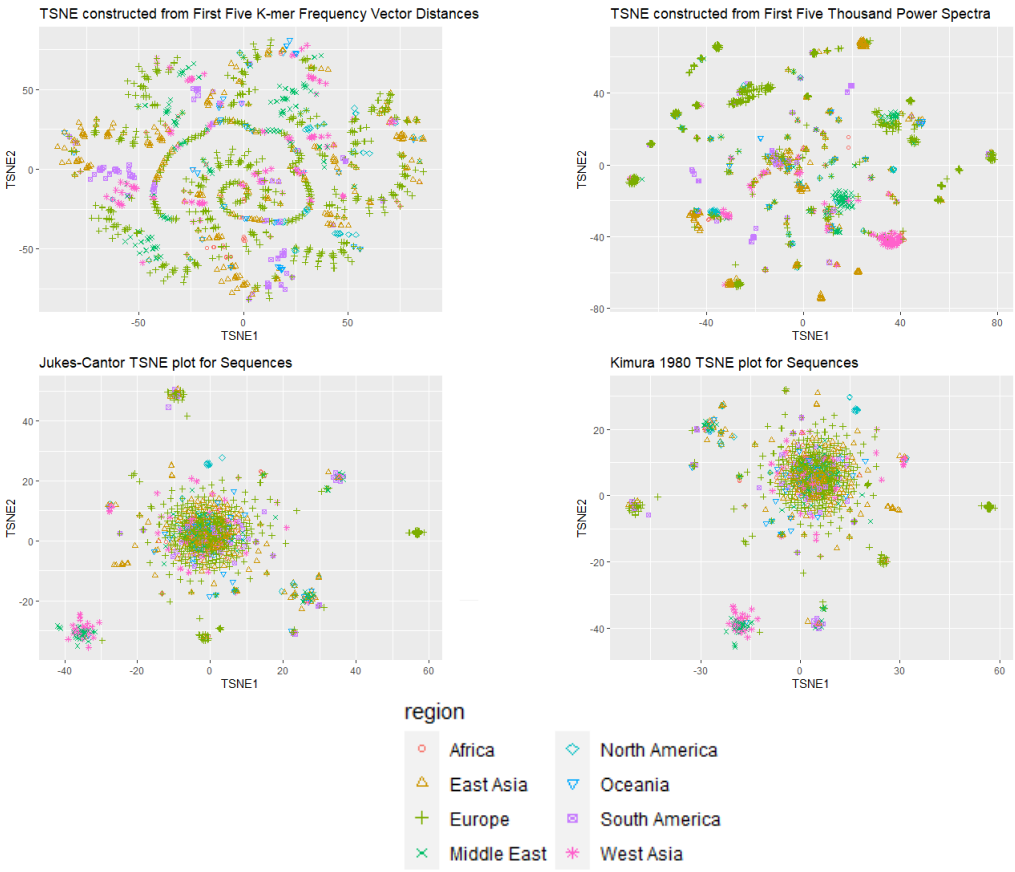


Figure 2.5: TSNE plots for different distance procedures

practical sense, as individuals from these two regions would be more likely to mix due to geographic proximity. The gold triangles representing East Asia are the most spread over the plot, indicating that individuals from this area perhaps carried the virus to other geographic regions.

The most prominent feature of the TSNE plots for the MSA-based methods is the large European cluster in the center of each plot. GISAID, based in the Federal Republic of Germany, expectedly obtained an imbalanced set of submitting regions far in favor of the proximal European countries, which explains the large cluster in the middle of the plot.

Another perhaps initially surprising result in Figure B.6 is the clearly patterned behaviour of the K-mer frequency vectors. This can be explained by redundancies which are present in the data (for example, given the listing of all five-mer frequencies alone (1024) one could

theoretically recompute that of the four-mer, three-mer, two-mer, and one-mer directly.

In total the Multiple Sequence Alignment (MSA) process using clustal-omega required 21 hours, 44 minutes, 52 seconds, which was partitioned among the three steps involved: Ktuple-distance (50 Minutes, 32 seconds), sub-cluster distance (16 Minutes, 14 seconds), and progressive alignment (20 hours, 38 minutes, 6 seconds). In terms of the computational time required for application of the distance assessment methods, the Jukes-Cantor and Kimura distance calculations are approximately 4000 times as computationally expensive (from a real time perspective) than the DFT distance. This is because the DFT distance does not require alignment, the Jukes-Cantor procedure uses multiple sequence alignment prior to computation of distances.

To count the K-mers in each of the 1,397 sequences and produce appropriate K-mer frequency vectors for comparison took 4,774 seconds. Scaling the frequency vectors vertically (across all elements in the sample) takes 0.16 seconds, and computing Euclidean distances takes approximately 2.15 seconds, for a total of 4,776 seconds. The power spectra calculation in total took only 392 seconds, which was partitioned among the process as follows: 2D Voss Encoding (19 seconds), Fourier Transform and Power Spectra Calculation (311 seconds), even scaling of ensemble (9 seconds), scaling PS vertically (5 seconds), and computation of Euclidean distances (49 seconds). Times for the alignment-free distance calculations were computed using the R package associated with this chapter's work [Thornton, 2021].

2.4.2. Multivariate Analysis of Variance

There are 30,563 Fourier coefficients for each genome. MANOVA is computed with the first 1000 Fourier Coefficients to provide a sufficient collection of points to extract patterns among the groups. The Fourier coefficients meet the criteria for normality, as the normality assumption is proven to hold asymptotically for Fourier spectra [Kawata, 1966].

For the MANOVA test, the null and alternative hypothesis are:

- H_0 : There is no difference in the population vector mean of the first 1,000 Fourier

Coefficients among the eight regionally submitting locations.

- H_1 : There is at least one pair of submitting locations for which the vector mean is different.

Therefore, a large statistic, and small p-value indicates that there is evidence that at least one pair of vector-means are different from at least one of the others. Table 2.3 shows that the p-values for this analysis are close to 0, indicating strong evidence for differences among the vector-means for the geographic regions. Results are shown for both K-mer counting and Fourier coefficients. The tabulated results are from Pilai's trace statistic for a MANOVA hypothesis test for the differences in population vector mean coefficients across the eight geographic regions in Table 2.1.

Table 2.3: MANOVA Results Table

Method	Pilai Statistic	Approximate F	Numerator DF	Denominator DF	p-value
K-mer, $k = 5$	2.17	5.84	700	9072	$< 2 \times 10^{-16}$
FC PS	6.49	5.05	7000	2772	$< 2 \times 10^{-16}$

An auxiliary procedure associated with MANOVA (Table 2.3) is the analysis of canonical variables. The coefficients for the canonical variables are the eigenvalues of the matrix product of the within and between matrices computed by the MANOVA procedure. From this perspective, the coefficients represent the linear combinations of variables which produce the largest separation between the levels of the factor of interest in the analysis.

Figure 2.6 displays a plot of the first two canonical variables after application of MANOVA with the power spectra from each sequence as the responses and the geographic region as the explanatory factor. The first two canonical variables following the analysis provide a very clear separation of the particular regions, with regions in geographic proximity consistently grouped together.

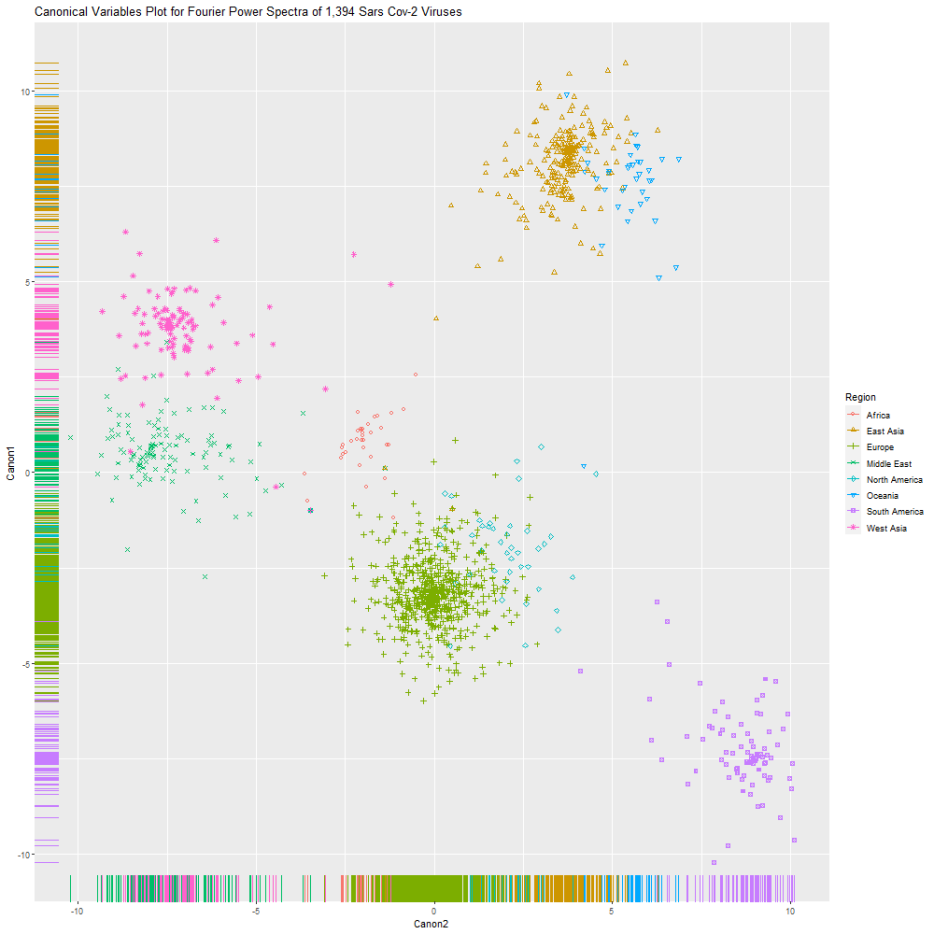


Figure 2.6: Fourier Coefficients Canonical Correlations Plot

2.4.3. Supervised Learning by Geographic Region

After the unsupervised approaches showed clusters of coefficients corresponding to approximately eight geographic regions, several supervised approaches were used to examine whether the Fourier spectra could be used to successfully classify the submitting region. First an Error-Correcting Output Codes model (ECOC) is used to train an ensemble of support vector machines to classify the location of data submission [Cortes and Vapnik, 1995]. The overall accuracy for this model was 71%, which was lower than any other method tested. This insight indicates that the Fourier spectra are not well partitioned linearly; therefore, several nonlinear models were applied to the classification. The first was random forest [Breiman, 2001], for which multiple settings displayed highly accurate classification results.

The first setting utilized 50 trees, and attained an accuracy of 79%, displaying clear superiority over ECOC SVM. This result corroborates that separation among the geographic regions is best done with a nonlinear classification method. When using a random forest of 500 trees, the accuracy increased to 80.5%. Additionally, we classified the data using neural networks [Hopfield, 1982], regression trees [Brieman et al., 1984], k nearest neighbors ($k = 10$), and naïve Bayes methods [Hastie et al., 2008]. Table 2.4 shows the full results.

Table 2.4: Various Classification methods for the Location of submission, on K-mer frequencies and Fourier Power Spectra.

Supervised Learner	**K-mer Vectors (Interval)**	**DFT Power Spectra (Interval)**
Naive Bayes	0.424 (0.411, 0.438)	**0.593 (0.580, 0.607)**
Regression Tree	0.179 (0.169, 0.189)	0.191 (0.181, 0.202)
K-Nearest Neighbors ($k = 10$)	0.722 (0.710, 0.734)	**0.776 (0.765, 0.787)**
Random Forest (500)	0.651 (0.639, 0.664)	**0.805 (0.795, 0.816)**
Neural Network (1 HL - 30 N)	0.505 (0.492, 0.519)	**0.580 (0.567, 0.593)**
SVM	0.688 (0.676, 0.700)	0.712 (0.699, 0.724)

The supervised learning approaches were implemented and validated using five fold cross validation to ensure that there was a representative proportion of each of the regions available in each of the training sets. The overall cross-validation accuracies for the K-mer frequency vectors and the Fourier Power Spectra as well as their 95% Binomial confidence intervals are displayed in Table 2.4. Table 2.4 contains a comparison of the use of the first-five K-mer frequency vector and the Fourier power spectra for the supervised learning and prediction of geographic location of origin for SARS-CoV-2 genomes. The key finding is that the Fourier power spectra outperforms K-mer frequencies in terms of classification for each of the learners trained. This is an encouraging result as it was faster to compute the DFT power spectra for the sequences than to count K-mers in a sequence.

The purpose of Table 2.4 is to show that the Fourier Power Spectra for these sequences can be used as valuable numerical summaries on which to train classification procedures that will subsequently predict the location of submission/origin for unknown samples that are submitted. As is evidenced by the bolded values in Table 2.4, the power spectra produced

more accurate results than did the K-mer vectors in nearly 70% of the cases. This key finding, while indicating a preference for the use of the power spectra for classification, should be taken with the understanding that the K-mer frequency vectors are actually much smaller, and hence contain far less information about the original sequence than do the Power Spectra.

2.4.4. Correlation between Distance Methods

The correlation between the Jukes-Cantor, Fourier PS distances, and several other post-MSA methods using Pearson's correlation is investigated to determine whether the Fourier power spectra distances contain ancillary or isomorphic information to the other distances. The results are in Figure 2.7. The matrix displays Pearson's correlation as computed among the distances that are produced via first five-kmer distances (fivemers), the Fourier Power Spectra distances (DFTPS), and the post-MSA methods that are provided by the `ape` package in R [Paradis and Schliep, 2019]. For a complete listing of the post-MSA methods, see the `ape` package documentation for the `dist.dna` function.

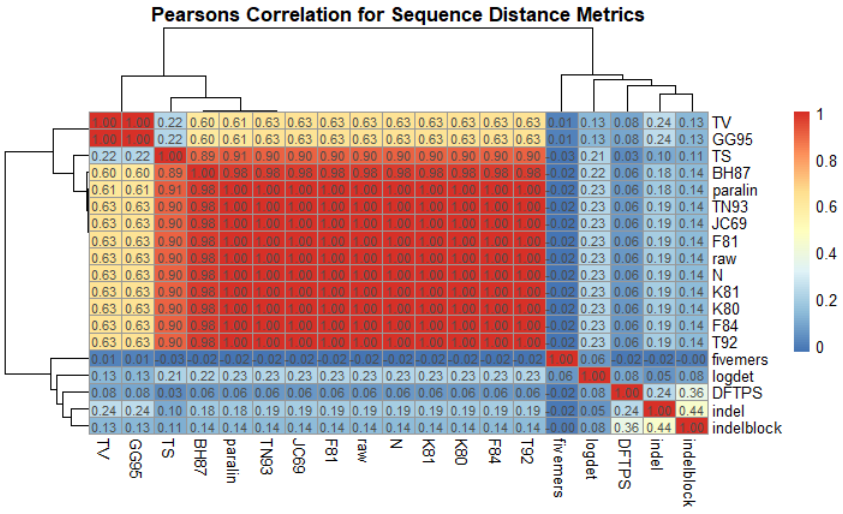


Figure 2.7: Correlation Matrix for Distance Methods

A strong correlation between any pair of the distances indicates that the respective distance calculation methods summarize similar information. As is clear from the matrix in Figure 2.7, there is not much correlation between the distances calculated from the Jukes-Cantor procedure and the Fourier PS distance procedure for the full length genomes of SARS-CoV-2. This indicates that the two distance metrics isolate different kinds of information from the sequences. Intuitively, this is expected as the Jukes-Cantor results are determined using an edit-distance like procedure, whereas the Fourier PS distances are computed using Euclidean distance among the Fourier power spectra.

The Jukes-Cantor method aligns all supplied sequences, and then determines the number of mutations among all sequences to calculate pairwise distances [Jukes et al., 1969]. The mild correlation found between the DFTPS distance metric and the indel/indelblock [Paradis and Schliep, 2019] methods indicates that these methods are retrieving at least somewhat isomorphic distance information. Both the indel/indelblock methods take a considerable amount of time to compute considering that they are post-MSA methods, but the DFTPS method is fast. In future work we will assess the ability of the DFTPS method to predict the distances that are computed by the indel/indelblock methods. The combination of information from these different distances for a total distance calculation will also be investigated. The procedures in Figure 2.7 are largely variations on the traditional substitution approach to sequence distance calculations [Strimmer et al., 2009]. A listing of the procedures compared in the correlation matrix is provided in Table A.1.

2.4.5. PS Filtering Results

In this section, the three filtering methods: MVF, AFL, and MVPCF are applied to a sample of viral genomes of the SARS-CoV-2 virus captured from the collection curated by the GISAID Initiative [Elbe and Buckland-Merrett, 2017]. The filters are first discussed individually and results are displayed regarding the unique nature of the approaches. This is followed by a comparison of the filter designs by two criteria.

First, the correlation of distances produced by the filter power spectra to the distances computed by the full power spectra is displayed as the filtered data size increases to the power spectra size. Second, the accuracy of the 500 tree random forest for computing the geographic location of origin for the data was assessed, and tabulated for display.

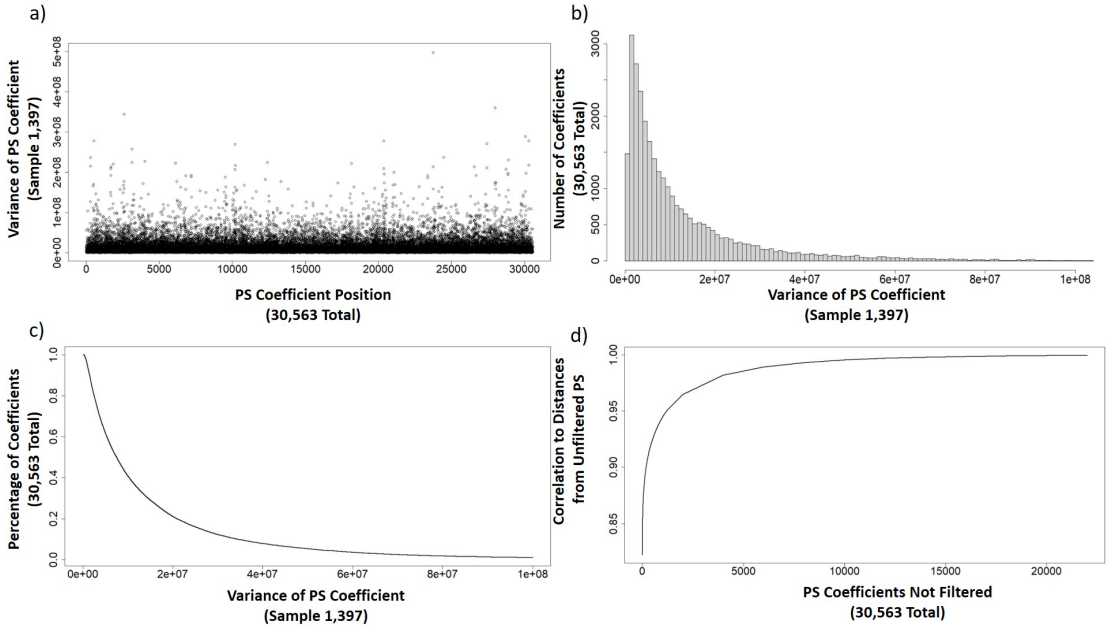


Figure 2.8: Variance Diagnostics for Power Spectra

First, MVF is applied to the set of PS for the virus genomes. Figure 2.8 shows several views of the PS component variances for the data to assist with visualization of the filter. Figure 2.8 is intended to depict a few of the graphical displays of PS data that one may consider when designing a MVF and selecting the η^* value.

Figure 2.8 (a) is a scatter plot of the sample ($N = 1,397$) variances PS Coefficient, η^* may be graphically visualized as a horizontal line, below which PS coefficients are discarded. A histogram of PS coefficient variances in Figure 2.8 (b), allows graphical visualization of η^* as a vertical line, left of which the area highlighted provides the number of coefficients filtered. Figure 2.8 (c) is a survival curve of coefficients by their variance, that is, the percentage of coefficients with variance greater than or equal to the ordinate is displayed as the abscissa. Figure 2.8 (d) displays the correlation of distances produced by the filtered and unfiltered

PS for a range of coefficients filtered, symbolically q.

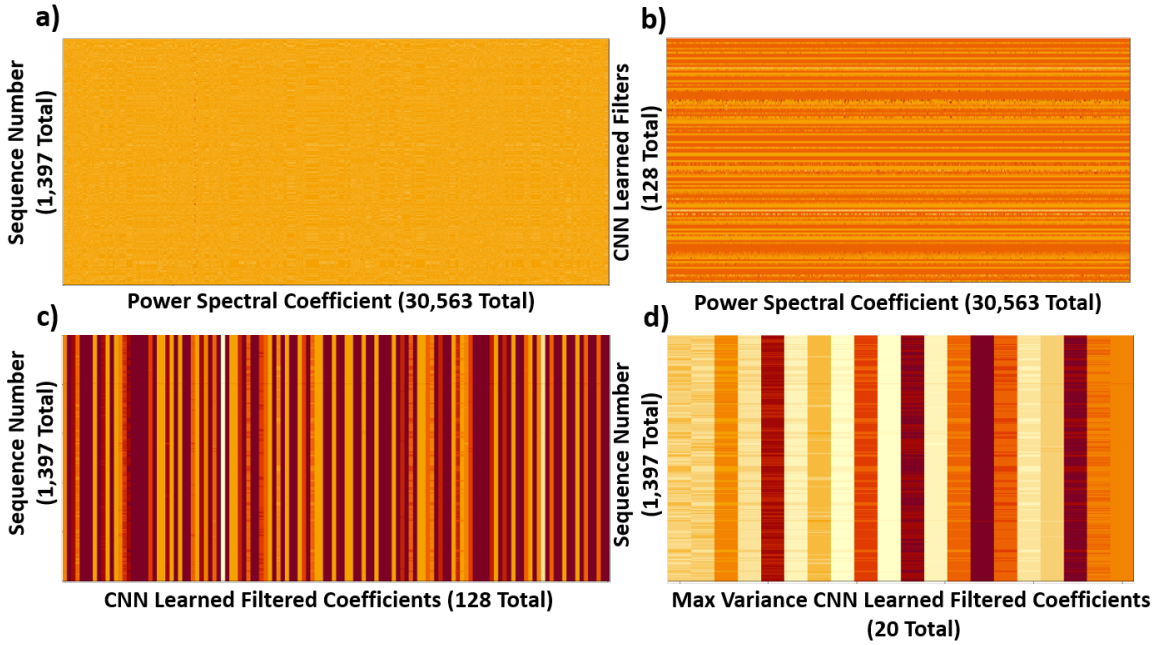


Figure 2.9: Some representative filters produced by the AFL method

A CNN for classifying the GISAID August SARS-CoV-2 PS by submission region for eight general regions extracted from the headers is implemented and trained using the `keras` package [Allaire and Chollet, 2021].

The network used 128 filters spanning the full PS on all 1,397 sequences. It was trained for 20 epochs, using a batch size of 100. A graphical display of the learned filters is provided in Figure 2.9. For the comparison section, the CNN trains different numbers of filters to provide a valid filtered PS set for comparison to the other two methods.

In MVPCF there is an inherent restriction on the maximum size to which the sequence may be reduced. Due to the nature of PCA, the number of components that can be produced from a set of sequences is limited by the number of sequences N. This N is the maximum number of PS coefficients that may be considered when producing the filter, so in the case of these 1,397 viral genomes, we are limited to considerations of the PCA of the maximum variance 1,397 PS coefficients. As stated before, when N is of greater length than m this is not of concern; however, in this case N is 1,397, and m is much larger ($m = 30,563$). Of

course, once the PCA of the 1,397 maximal variance coefficients is performed, selection of the first k components allows for an even greater reduction.

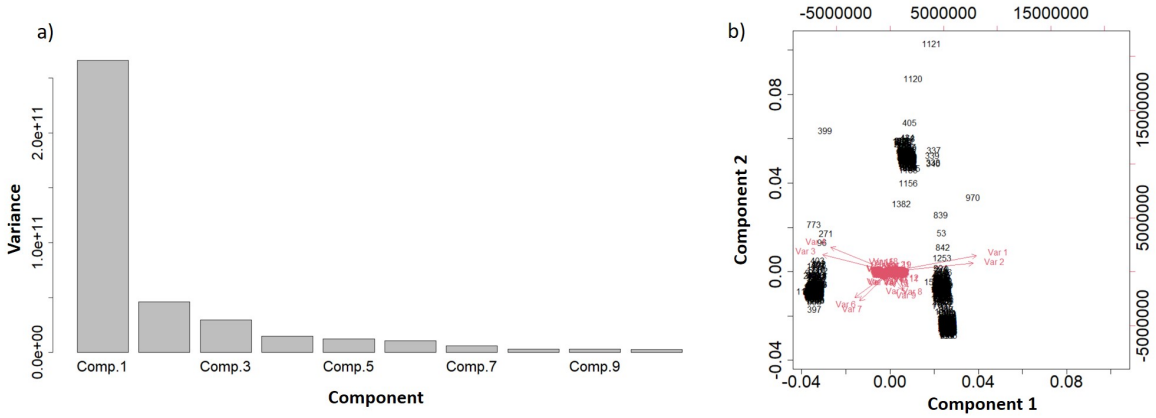


Figure 2.10: Scree Plot and Bi-Plot

Figure 2.10 provides visualization of the resulting PCA for the maximal variance 1,397 elements of the PS. The screeplot in Figure 2.10 (a) shows that the vast majority of the variance of the set of maximal variance PS coefficients is contained in the first component. Distances calculated from this single value are correlated with distances calculated from the entire PS with $\rho \approx 0.88$. The biplot in Figure 2.10 (b) shows that even the first two coefficients provide a large separation of the data, notably three pronounced clusters appear.

MVPCF produces Euclidean distances which are most linearly related to the Euclidean distances computed from the full PS for the same number of coefficients in the other two methods, as per Figure 2.11. That said, recall that MVPCF has the ability to consider a maximum of only N elements from the unfiltered PS (1,397 in this case). Therefore, there must be some $N^* > N$ such that the MVF procedure will outperform the MVPCF's best possible correlation, which occurs at N. Perhaps one of the more surprising aspects of Figure 2.11 is the low correlation from the CNN trained filters. These filters extract information more relevant to differentiating the region of submission than correlating to the full PS distances. Hence AFL is not the best choice of the three methods for correlating to distances from the full PS. When it is desired that general information only is considered

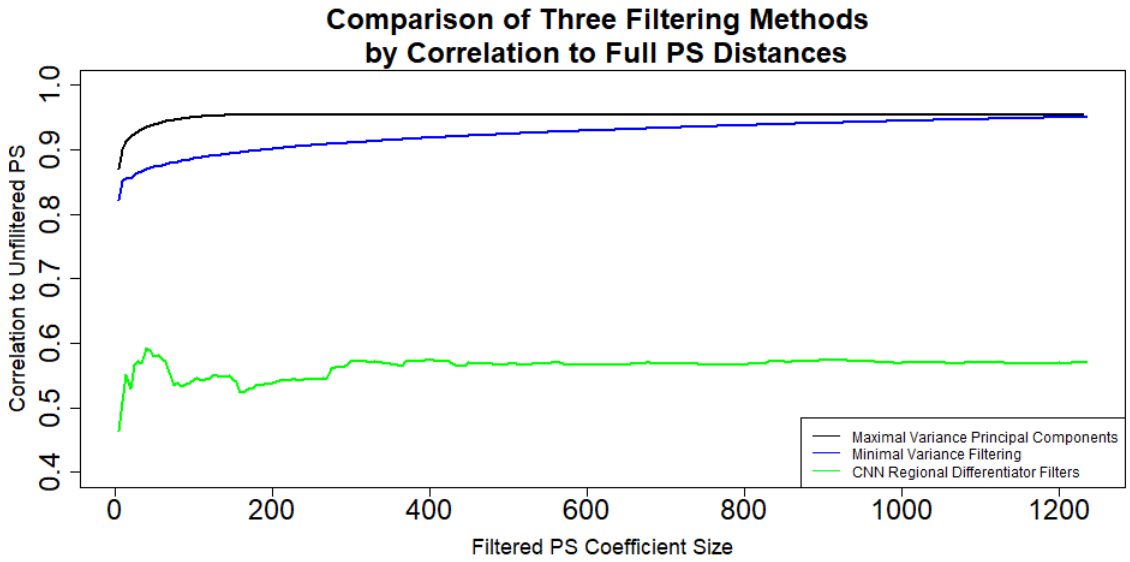


Figure 2.11: Filtering Methods Comparisons, by Correlation to Full PS Distances

in differentiating the sequences, the AFL procedure should use labels which indicate general distance classes.

Table 2.5: Random Forest of 500 trees CV Accuracies for Region Classification from Filtered PS (%)

Filtered PS Size	**MVF**	**AFL**	**MVPCF**
50	24.833	**45.239**	21.759
100	24.917	**47.740**	38.874
250	31.003	45.528	**48.245**
500	28.706	48.244	**48.532**
1000	45.451	44.516	**48.532**

To compare filter efficiency at capturing specific important information from the PS, five fold cross validation accuracy (in percent) of regional classification for the SARS-CoV-2 data with a random forest classifier (500 trees). Specifically, Africa has 35 sequences, East Asia has 257, Europe has 678, Middle East has 153, North America has 42, Oceania has 38, South America has 89, and West Asia has 105. are displayed in table 2.5. As we can see, the AFL method extracts the best information for discerning the region of submission lower sized filtered PS. This was expected because these labels were used in training the filters. However, for larger filter sizes, the MVPCF method actually produces a higher accuracy.

2.5. Conclusions

The results of this experiment show that there is some merit in using the sequence summaries provided by the Fourier coefficients to determine a rough estimate of the region of submission, and by extension, to other classes. The distance information captured by the Fourier coefficients appears to be uncorrelated with the distance information captured by the Jukes-Cantor procedure, and the K-mer frequency distances, indicating that the three extract different (but valid) information from the sequences.

The Walsh coefficient simulation experiment also indicated that distances from Walsh to Fourier coefficients for the genomic signals decrease to zero as the signal length increases. Although genomes are not technically smoothly varying functions, it is still valid to model them as such for the purpose of information extraction (at least in terms of the region of submission). A statistical hypothesis test for determining whether the data come from one group or multiple groups based on the Fourier coefficients is derived and discussed. Relatively small subsets of the initial power spectra may be used to capture approximately 80% of the same orderings of the pair-wise sequence distances from the full power spectra. Three filtering methods for determining the optimal coefficients were discussed. AFL can learn which information is most important in the full PS for classifying a specific label, and produce filters which generate values based on linear combinations of those elements.

This work is valuable as it displays the ability of the Fourier coefficients for summarizing certain vital aspects of the information that is contained in the more lengthy sequence. Further, a smaller subset of the Fourier coefficients that are produced by a single sequence might be used to uniquely identify that sequence, and give accurate information about the true distances between sequences. The classification analysis indicates that the Fourier coefficients may be used to determine the general submitting region for a particular sequence with reasonable accuracy, indicating the potential use for these techniques in determining particular locations on potentially smaller scales than are undertaken in this work. The study using the SCV2 genomes shows that the techniques here are applicable to viral genomes as

well as those in animals and bacteria, and may be used to estimate the region specific genetic drift of viruses especially during outbreak events like the SCV2 Pandemic of 2020.

There are a few motivating directions in this study, however one of the first avenues to be addressed is the implementation and evaluation of the statistical hypothesis test derived in this chapter. Due to the nature of the test, it may be used for any comparison of ensembles of Fourier coefficients, not just genomic Fourier coefficients. As such, the test is general, and it should be assessed by application on various different kinds of harmonic datasets. An overall analysis of the robustness of the procedure to violations of the second-order stationarity assumption could also shed some light on its utility for comparing genomic sequences.

In this work, primarily the power spectra of the Fourier series is examined; however, a substantive amount of organizational material is contained also within the phase spectra. As such, a comparison of the classification capabilities of the two components, and their combined capabilities should be investigated. The result of a study of this kind might indicate the preferential use of one part over another or mixed portions of the two under special circumstances.

The adaptation of the approach for examining protein sequences may also be of interest, as well as the ability of the approach to be used on other kinds of biological signals such as methylation proportions, protein concentration levels, or other methods for the differentation of organisms.

Chapter 3

Analysis of Compositional Data in Antibody Glycosylation

Tuberculosis is a deadly infectious disease that is caused by the bacillus *Mycobacterium tuberculosis*. Tuberculosis, as recently as 2020, is the leading cause of death due to an infectious agent [Global Tuberculosis Report 2020, 2020]. The disease usually occurs in one of two cases, active and latent. Latent tuberculosis patients do not actively spread the infection to others, while active cases spread the bacillus via respiratory processes. The antibodies produced in response to foreign pathogens play a role in the body's defense against disease. The creation and release of specific kinds of antibodies is a key component of the human immune response to infectious diseases like tuberculosis [Seeling et al., 2017] [Gudelj et al., 2018] [Gunn and Alter, 2016] [Lux and Nimmerjahn, 2011].

A major part of the vertebrate immune response is the usage of antibodies for the detection and neutralization of infectious agents. Due to the enormous number of potential pathogens, the form of the antibodies must be highly variable and allow for frequent modifications. Sometimes the antibodies are modified after being translated directly from the genetic sequences. These kinds of mutations, known as somatic mutations [Tonegawa, 1983], are a part of the adaptive immune response which allows for the wide variety of antibodies in response to various kinds of pathogenic agents [Braden and Poljak, 1995]. Another such modification is the attachment of small multi-sugar chains, called glycans, to different positions on the antibody. Glycans are considered to come in different *species* depending on the type of sugars that make up the chains [Alberts et al., 2002]. In addition to the somatic mutations which contribute to the wide variety of potential antibodies, mechanisms for *post-translational* modification are also present.

One specific class of antibody known as Immunoglobulin G (IgG), which is the most common kind of antibody, is particularly amenable to study. the IgG antibody molecules

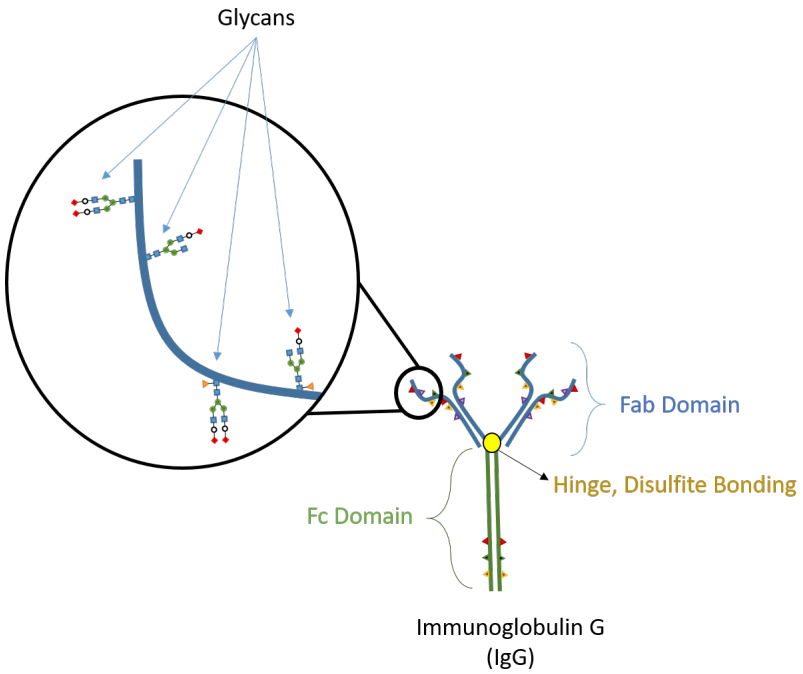


Figure 3.1: Structure of IgG for reference

produced will target bacteria like *Mycobaterium tuberculosis*, which is the cause of tuberculosis. It is further known that the addition of oligosaccharides, known as glycans, in specific locations along the IgG molecules influence the ability and affinity of the IgG molecule to bind to the pathogen as well as the immune response cells necessary for detection and elimination of the pathogen [Alberts et al., 2002].

The IgG molecule itself is composed of two light peptide chains and two heavy peptide chains that are bonded together via disuflite bonding. This is typically depicted in a cartoon form as a "Y"-shaped molecule, where the two light chains are attached along the split portion of the "Y", and the heavy chains run the length of the entire body of the "Y", as depicted in Figure 3.1. The antigen binding sites are at the ends of the two arms of the "Y" are generally identical. The arms of the "Y" are labeled the Fab portion of the IgG molecule, and the tail is known as the Fc region. The Fab domain of the IgG molecule binds to antigens, and the Fc domain binds to the Fc receptors of human immune cells which allow for the proper handling of detected antigens [Alberts et al., 2002]. IgG is characterized by the relative percentages of

glycan species that are present on the Fab and Fc subdomains of the molecule. The different glycosylation patterns of the antibodies provide information about the immune response of the patient from which they were sampled [Alberts et al., 2002].

Scientists are able to quantify the relative proportions of glycan species in samples by applying an experimental procedure known as capillary electrophoresis. In this approach, the weight and electrical charge of the different glycan species causes them to propagate at different rates through a capillary with an induced electrical field. By examining the relative intensity of observations at the end of the capillary over time, a measure of the proportions of each glycan in the sample is achieved [Alberts et al., 2002]. A quantification of the glycans present in each of the subdomains of the IgG molecule may be made by inspecting a curve of detections over unit time, and extracting the area under the intensity curve surrounding known reported times [Mittermayr et al., 2013].

Linkages between *post-translational modification* and the outcome of disease types are usually of interest to researchers. Specifically, the relative abundances of glycan species has been shown to be useful in classifying tuberculous patients as either being "Latent", "Active", or "Negative" [Lu et al., 2020]. A presence of diabetes melletus, a known comorbidity often associated with the outcomes of tuberculosis, is also known for the subset of patients. A partial least squares discriminant analysis (PLS-DA) model was used to demonstrate that the relative proportions of these glycans can be used to differentiate the disease subtype of tuberculosis patients [Lu et al., 2020].

In addition to glycan concentrations, other variables of interest are quantified using various laboratory procedures that determine protein concentrations from peripheral blood samples. These are referred to as functional assays. The approaches discussed in this chapter involve combining information from the relative glycan abundances with the additional protein information to provide accurate predictions of the disease type. Specifically, we seek to predict the tuberculosis and diabetes status by investigating the glycan profiles and the protein information. This chapter is organized as follows: first, a description of the Partial Least Squares Regression and Discriminant Analysis approach (PLS-DA/R) is provided as

a gold standard against which to compare the new classification model prediction accuracy. This is immediately followed by the results of applying the PLS-DA procedure to a new set of glycan data gathered on tuberculous patients. Two new models, a semi-parametric model and a rank probability model, are used to differentiate disease subtypes are described. The semi-parametric model is introduced and defined for general compositional data, followed by application to the tuberculosis data. Next, the glycan rank probability model is introduced and the results for the tuberculous data are shown. These two models can be thought of as following the supervised learning model, where a training phase is followed by a classification phase. It may be more accurate to consider these two phases the a model-building and a prediction phase. Specifically, the relative IgG glycosylation proportions are investigated in their ability to separate tuberculosis patients into active and latent sub-classifications. The chapter closes with a few concluding remarks and future directions of this study.

3.1. Glycosylation on Antibodies in Disease Response

When electrophoresis is applied to the glycans attached to the antibody immunoglobulin G (IgG), it might be applied across the entire IgG molecule (whole), or within one of the sub-domains (Fab or Fc). In the case of the data discussed here, the whole samples are distinct from the Fab and Fc ones. In other words, The Fc and Fab glycosylation values refer to different sets of compositional data. Compositional data cannot be considered as independent observations because the measurements are taken relative to one another. Proportions or percentages that sum to a constant are one example of such data [Filzmoser et al., 2018a]. Occasionally, some of these compositional groups are missing or irretrievable for patients, leading to a need for multiple imputation. The KNN Impute procedure was previously used to impute missing compositions [Hron et al., 2010][Lu et al., 2020]. The KNN Impute procedure selects the group of k neighbors nearest to that missing, based on a Euclidean distance computed between all variables not missing, then imputes the median as the missing value.

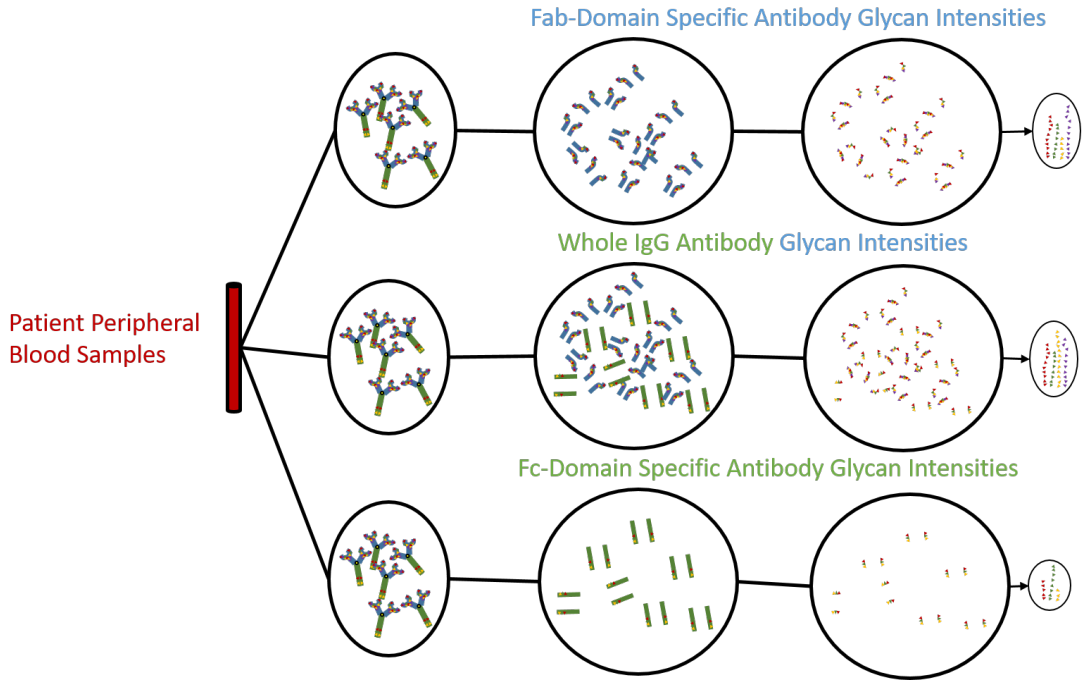


Figure 3.2: Example Data-Collection Multiple-Composition Creation

3.2. Data Description

Figure 3.2 is a simplified pictoral display of the breakdown of the sampled materials prior to electrophoresis and quantification. The high variability of the IgG molecules allows them to detect many different kinds of antigens. One factor that contributes to the variability of the IgG molecules is the wide variety of combinations of glycan species. For reference, a table of the different glycan species that are detected in the capillary electrohoresis experiment is given in Figure 3.3.

Unfortunately, certain species of glycans are indiscernible. In Figure 3.3, the species which are labeled with two alphanumeric strings separated by an underscore are species that cannot be resolved as a specific glycan, but are non-distinct. Specifically, "G1F_G1FB" refers to a peak where G1F and G1FB glycans cannot be differentiated, "G1FB_G2" refers to one where G1FB and G2 cannot be differentiated, "G1_G0FB", G1 and G0FB, and "G1_2" G1 and G2. These are reported as is in the tuberculosis data.

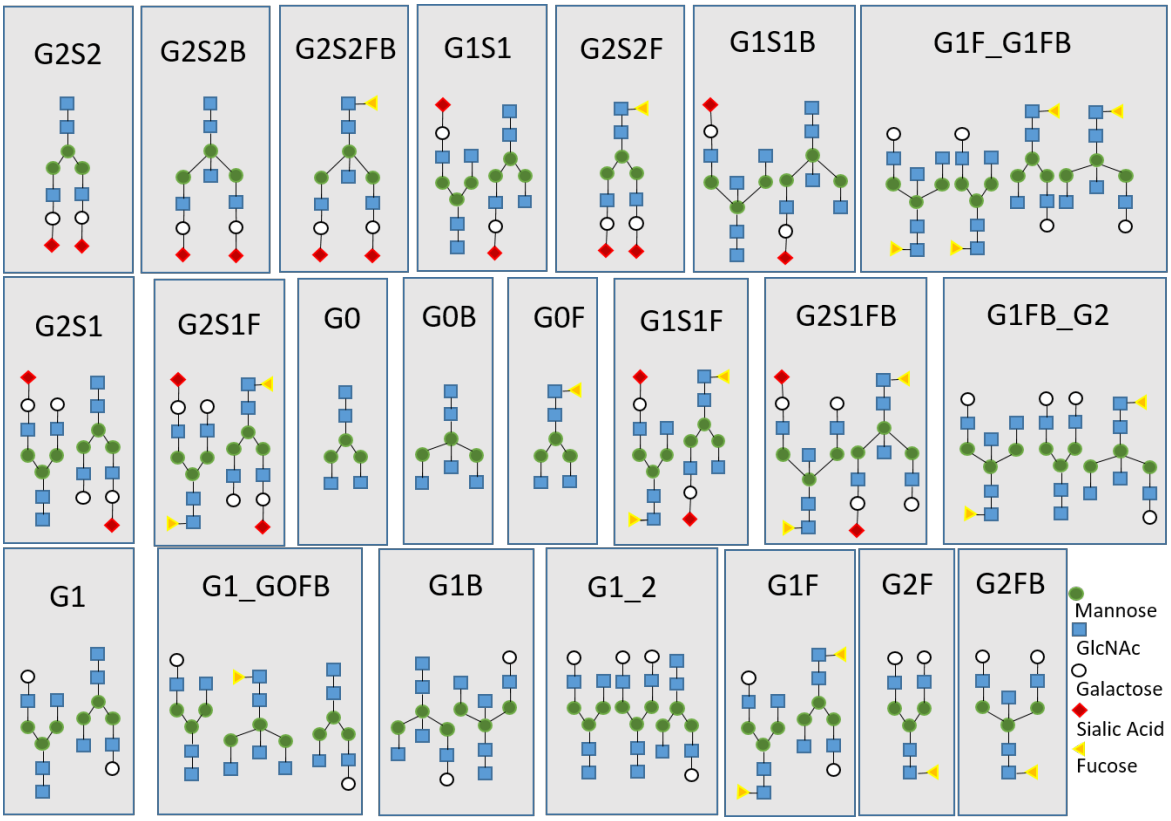


Figure 3.3: Types of Glycan Structures

This complication provides a very interesting question for future analysis. In compositional data, when there are vague class calls, what sorts of effects does this have on the actual estimates? This concept is further expounded upon in the future directions section of this chapter. In the present work, however, these categories were simply treated as their own distinct category, as if they were their own unique glycan species.

In addition to the glycan composition data, several other functional assays were performed, and the data provided for the Tuberculosis patients B.1. The non-glycan data listed in provides valuable information for the differentiation of Tuberculosis cases. Demographic information for the patients was also recorded, and is displayed in Table B.2. Due to the varied nature of the data, one potential modeling approach is to use an empirically determined distribution function to provide likelihood information for classification. Using a previously discussed methodology for collection [Lu et al., 2016], and the Applied Biosystems Hitachi

3500 xL Genetic analyzer for capillary electrophoresis [Berosik et al., 2010], the 2021 tuberculosis dataset [Lu et al., 2021] was captured for analysis.

3.3. Modeling Approaches

The remainder of this chapter deals with the modeling of the relative concentrations of glycans and other proteins. First, a classical approach for the modeling of related matrices of data, known as Partial Least Squares Discriminant Analysis (PLSDA), is provided, as was done with similar data [Lu et al., 2020]. Results of applying this procedure to the new antibody glycosylation data set are described.

Following this, a description of a semi-parametric likelihood model which treats glycan information in terms of it's compositional nature is described. The semi-parametric model forms log- likelihood products from multinomial distributions for the compositional portion of the dataset. The modeling capabilities are augmented with the observed empirical distributions of the demographic and functional variables. The model is first introduced theoretically, and then the results of an application of the model to prediction of disease subtype are presented and described.

Lastly, an approach to modeling the transitional nature between glycans is presented in the context of a Markov type model. This approach, named the glycan rank probability classification model, is again developed and results of application to the present dataset using only the glycan data. Some final considerations regarding a few other kinds of modeling techniques are discussed, and the efficiency of each of the models suggested with regards to the modeling of the IgG glycosylation dataset are offered in conclusion of this chapter. A few directions of investigation are provided as future works. The appendix associated with this chapter (Appendix B) contains a table of the noncompositional variables available in Table B.1, displays their empirical distributions functions within each tuberculosis/diabetes group in Figure B.1, and provides a table of the relevant demographic data from the set in Table B.2.

3.4. Partial Least Squares Modeling

As a reference point for comparison with the newly introduced models in this chapter, the PLS-DA procedure is described. This is one of the primary procedures that was used to link the glycan proportions with outcomes of interest [Lu et al., 2020] PLS-DA may be thought of as Principal Components Analysis (PCA) in two data-blocks followed by regression of one block of scores on another. As such, and particularly as noted in [Brereton and Lloyd, 2014], PLS-DA is useful as a descriptive model for determining which factors are the most explanatory of the separation between the classes.

The partial least squares technique can be used to predict numerical outcome variables. An extension of the regression procedure allows it to be used for the prediction of categorical outcomes. Outcome categories are encoded as a series of variables which take on two values then the partial least squares regression model is fit, then used to produce scores for each possible category for any new data, the class is selected by discriminating on the scores produced for each possible category [Geladi, 1988].

The method has been used extensively in the food sciences and other chemometric analysis experiments. One instance of PLS-DA use in industry from the food sciences concerns the review of sensory variables for food items. These variables often require the contracting of judges, or other measurement techniques which are relatively more costly than measurement of a few less expensive alternative proxy measures. Direct measures are quite expensive, yet the resulting data is quite valuable in terms of product quality assessments. Instead of paying to have these more expensive variables assessed directly, the estimation of these variables from sets of variables that are cheaper to collect and more readily available is one potential product of the PLS-DA model. This is also why a significant amount of the PLS-DA regression and discriminant analysis occur in experiments where a subset of the variables to be predicted are costly to determine. PLS-DA has also been recently shown to be useful in particular genetic experiments as well [Strimmer et al., 2007]. This potential application of the PLS-DA model presents another interesting usage question: can the PLS-DA model

be implemented in such a way as to *impute* missing observations? In other words, building the model with complete cases, and predicting the incomplete cases.

For more historical notes on the PLS-DA method, see the summary provided by Paul Geldi [Geladi, 1988]. In this work, Partial least squares analysis is applied to the 2021 tuberculosis data set, with a specific emphasis on the discriminatory ability of the procedure to determine the disease outcome categories. The approach of using the partial-least squares methodology for modeling these kinds of tuberculosis datasets is inspired by a similar tuberculosis dataset [Lu et al., 2020]. The previous work described the application of PLS-DA on a subset of the glycosylation values which were first selected using the Least Absolute Shrinkage and Selection Operator (LASSO) technique of Hastie and Tibshirani [Tibshirani, 1996, Tibshirani, 1997]. The glycosylation values were then subjected to the partial least squares regression technique primarily for visualization of the latent variables (LVs) that were produced.

In PLS-DA, latent variables are similar to principal components [Hotelling, 1933] with one extension; while simultaneously maximizing the the variance within the independent variables selected, the LVs are selected such that they have the highest correlation with a corresponding set of scores for the dependent variables. The PLS-DA procedure itself has been widely applied to many different experimental settings since it was introduced by Herman Wold in 1966 [Wold, 1966]. In the initial paper, PLS-DA was used as a post-hoc procedure, where the variables included in the model were those which were first selected using the LASSO approach. A later algorithm, implemented in MatLab, was a constituent part of a chemometric analysis tool suite called Chemoface [Nunes et al., 2012]. As a part of the reapplication of PLS-DA to a new tuberculosis dataset, the Matlab procedure was updated, and implemented in an R package named `PLSDANunes`.

From a broad perspective, PLS-DA and PCA are very similar in that they use measured variables to attempt to determine deeper trends that can be described by unmeasured variables. In a typical PCA, the principal components of the data are determined and displayed as a series of loadings which may be thought of as the relative importance of each of the vari-

ables in describing the underlying variables. After these loadings are determined through the analysis, the typical process is to attempt to use some expert or domain knowledge in conjunction with human interpretation to determine a meaningful underlying variable constituted by each observed component. In the case of PCA the underlying variables are mathematically orthogonal to one another; furthermore, they are ordered in terms of the variable combinations that explain the most variability in the response. The key difference when considering PLS-DA is that the variables, while still maximizing the variance, are also chosen to produce the largest correlation with the response variables.

A few notes about PLS-DA overall, and about it's particular application in this case that should be made:

1. PLS-DA, like PCA, is most powerful when the underlying interdependent relationships among the independent and dependent variables are linear. In other words, when the relationship between the independent underlying variables and the dependent underlying variables is linear. This is made clear by the fact that the procedure overall attempts to maximize linear measures of correlation between the two.

2. PLS-DA has had significant confounding associated with its usage in the past; to the point that the original acronym has been widely accepted as having a dualistic interpretation. That is, PLS itself can be interpreted as either Partial Least Squares (which was the initial intention of Herman Wold - the inventor) *and* Projection to Latent Space/structures, the arguably more descriptive of the two names. This is unfortunate, while both meanings hold some value in the overall interpretation of the technique, it complicates a complete understanding of the method to have two different names floating around for the same procedure.

3. PLS-DA often produces *latent variables* which are not as powerful as principal components determined from the same data[Ruiz-Perez et al., 2020].

4. PLS-DA is often applied as a simple supervised learning technique without great consideration as to the modeling assumptions that are implied by its usage. These neglected

assumptions sometimes create an interpretive *black hole* where the effects have little or no practical meaning.

5. Using PLS-DA as a method for selection of variables in subsequent linear modeling approaches suffers from the curse of multiple hypothesis testing, and this is rarely corrected.

In regards to the initial tuberculosis dataset modeled using PLS-DA [Lu et al., 2020], the glycans exhibit a compositional relationship, especially in that they have already been normalized prior to modeling. There have been a few suggestions for manipulating compositional data into a form for which it is more appropriate for use with the PLS framework [Wang et al., 2010]. These usually involve some transformation of the data such as the *centered log-ratio transform* [Aitchison, 1982].

Recall that PCA a multivariate procedure that can be applied to observations of multivariate data $\boldsymbol{X}$. In this case $\boldsymbol{X}$ represents a matrix of n observations (rows) of p distinct variables (columns). Let x_{ij} refer to the i^{th} observation of the j^{th} variable. Furthermore let the row vector $\boldsymbol{x_{i\cdot}}$ refer to the i^{th} row of the data matrix $\boldsymbol{X}$ and $\boldsymbol{x_{\cdot j}}$ refer to the j^{th} column of the data matrix $\boldsymbol{X}$.

The decomposition of the data matrix $\boldsymbol{X}$ into its principal components can be expressed as the problem of determining the singular value decomposition of the variance-covariance matrix of $\boldsymbol{X}$. Assume that the data matrix $\boldsymbol{X}$ has already been normalized and scaled, such that $\frac{\sum_{i=1}^{n} x_{ij}}{n} = \mu_j = 0$ and that $\frac{\sum_{i=1}^{n}(x_{ij}-\mu_j)^2}{n-1} = \sigma_j^2 = 1$. Note that the variance of the j^{th} column of the data matrix may be written simply as $\sigma_j^2 = \frac{\sum_{i=1}^{n}(x_{ij}-0)^2}{n-1} = \frac{\sum_{i=1}^{n} x_{ij}^2}{n-1}$. Furthermore, the covariance of the j^{th} and k^{th} variable can be expressed as $\sigma_{jk} = \frac{\sum_{i=1}^{n}(x_{ij}-0)(x_{ik}-0)}{n} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{n}$. This means that the variance-covariance matrix for the full data matrix $\boldsymbol{X}$ may be written in matrix form as $n^{-1}\boldsymbol{X}\boldsymbol{X}^T$.

A complete formal derivation of PCA is provided by Anderson [**?**]. A derivation of PCA is provided in this context as a key progenitor of the PLS procedure. The derivation amounts to the enumeration of a few key realizations:

- For p-vector random variable $\boldsymbol{X}$, assume $E(\boldsymbol{X}) = \boldsymbol{\mu_X} = \boldsymbol{0}$, or transform it such that

this is the case.

- An arbitrary linear combination of the random variables in $\boldsymbol{X}$ in terms of scalars $\beta_1, \beta_2, \ldots, \beta_p$; that is, $(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$, can be written for non-random variable p-vector (column vector) $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^T \boldsymbol{X}$.
- The expected value of the linear combination is $E(\boldsymbol{\beta}^T \boldsymbol{X}) = \boldsymbol{\beta}^T E(\boldsymbol{X}) = 0$.
- The square of the expected value of the linear combination $\boldsymbol{\beta^T X}$, $E(\boldsymbol{\beta^T X})^2$ is 0.
- The expected value of the square of the linear combination, $E\left((\boldsymbol{\beta^T X})(\boldsymbol{\beta^T X})^T\right)$, can be written as $E\left(\boldsymbol{\beta^T X X^T \beta}\right)$.
- The variance of the linear combination, $\mathrm{Var}(\beta^T X)$ is $E\left((\boldsymbol{\beta^T X X^T \beta})\right)$.
- Since $\boldsymbol{\mu_X} = \boldsymbol{0}$, let $\boldsymbol{\Sigma} = E(XX^T)$ represent the variance-covariance matrix of the p-random variables $X_1, X_2, \ldots, X_p$.
- To ensure that the vector $\boldsymbol{\beta}$ is normalized constrain $\boldsymbol{\beta}$ such that $\boldsymbol{\beta^T \beta} = 1$.
- performing constrained optimization in the usual way with a Lagrange multiplier λ, while enforcing the above constraint, $(\boldsymbol{\beta^T \beta} = 1 \iff \boldsymbol{\beta^T \beta} - 1 = 0)$, the maximizing $\boldsymbol{\beta}$ of the variance under this constraint is the value of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, such that $\phi = E\left((\boldsymbol{\beta^T X X^T \beta})\right) - \lambda(\boldsymbol{\beta^T \beta} - 1)$ is maximized.
- The vector of partial derivatives of the above expression for each β_i are given by $\frac{\boldsymbol{\partial \phi}}{\boldsymbol{\partial \beta_i}} = 2\boldsymbol{\Sigma \beta} - 2\lambda \boldsymbol{\beta}$.
- Setting each of the partial derivatives equal to zero produces the equations $(\boldsymbol{\Sigma} - \lambda \boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{0}$.
- From which it may be noted that to enforce $\boldsymbol{\beta}^T \boldsymbol{\beta} = 1$, the familiar characteristic equation for the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ must hold, $|\boldsymbol{\Sigma} - \lambda \boldsymbol{I}| = 0$.
- Therefore, it is the case that determining the eigenvectors of the variance covariance matrix will provide the estimates of β with the maximum variance, and constrained such that the inner product is one.

In PLS, two data matrices, the independent/dependent, explanatory/response, or factor/class, denoted $\boldsymbol{X}$ of size $n \times k$, and $\boldsymbol{Y}$ of size $n \times m$ are related. PLS seeks to maximize the variance of the components in both $\boldsymbol{X}$ and$\boldsymbol{Y}$. A third criteria is sought in PLS which is conceptually not present in PCA, but is in Principal components regression. That is, a determination of the linear relationship between the latent variables from the dependent data, and the latent variables from the independent data [Wold, 1966]. In theory this amounts to a determination of the principal components of the $\boldsymbol{X}$ and the $\boldsymbol{Y}$ data, and linking the two determined latent structures together in an ordinary least squares sense.

The discriminant analysis portion of PLS refers mostly to encoding a class membership indicator matrix from the classification and using this as the dependent block of variables in the analysis. The estimated latent variables can then be used to determine the associated class membership for new data, or for the data used to train the model. The PLS method cannot achieve perfect prediction in the set of data used to train it if fewer coefficients are examined than data points $A < n$. A PLS and PCA type study of the tuberculosis data set shows that this modeling procedure allows for some of the case differentiation of interest.

3.4.1. Data Study: 2021 Tuberculosis Data - PLSDA

The PLS procedure was applied to each of the different glycan composition sets for the 2021 tuberculosis dataset separately. It was also applied to the functional assay data, as described in Table B.1, as well as the combination of the three glycan compositions and the functional assay data.

The first outcome was a class membership matrix, which had five categories. The categories are active tuberculosis with and without diabetes, latent tuberculosis with and without diabetes, and negative for tuberculosis and diabetes. The resulting data visualization from this section is included in Appendix B.

Plots of the first two components of the model for a few different $\boldsymbol{Y}$ class membership blocks are shown. The first response classification considered was the patient's combined

tuberculosis and diabetes status, which had five groups as described above. The $\boldsymbol{Y}$ matrix contained five columns in this case. The resulting plots of the scores of the first and second latent variables (LV) are shown in Figure B.2.

Recall that the LVs computed in an analysis with the PLS procedure depend on the class data. This means that the LV plots shown for the five different groupings of variables in Figure B.2 are different from the LV plots shown for only three groupings, such as tuberculosis status alone which is shown in Figure B.3.

The figures in B.4 show how the data scores align on the first two LVs for the diabetes classification data. These differences highlight why partial least squares might sometimes be considered a supervised PCA, where the user supplies the labels of the data in addition to the input during the training phase, and also selects the number of LVs to produce.

Principal components plots for the same data for the first two principal components were also produced and colored according to the five group separation which is shown in Figure B.5. The principal components do not depend on the classification data, and are the same no matter which method is chosen; therefore, the legend for the plot contains Tuberculosis/Diabetes class labels.

It is sometimes also helpful to examine a TSNE plot (see 2.3), which allows for display of high dimensional data using two or three axes. Figure B.6 shows TSNE plots created using different variable subsets, the plot shows a minimal degree of separation of the clusters in this case.

PLS-DA can produce a model based on a user-defined latent space dimension. In practice this is accomplished by calculating only the first A coefficients as requested by the user. This is possible as long as the residual matrix remains computationally distinguishable from zero. Figure B.7 shows the accuracy in PLS-DA for differentiating the proper joint tuberculosis/diabetes when trained using different subsets of the variables. The accuracy within each class varies depending on the number of latent variables used in the analysis, but in every case the more components used the higher the accuracy.

In summary, the PLS-DA method can be used for multiclassification problems and help

to determine linear combinations of variables that may best relate a set of observations to their corresponding classes. It also allows for some degree of variable selection. In a similar way to principal components analysis, a key part of the procedure remains in finding a way to interpret the loadings produced by the model.

3.5. Semi-Parametric Modeling

From a natural point of view, the ratios of kinds of glycans in a sample at a given point in time may be thought of as a set of samples taken from a fixed total n where each sample fits into a single category.

Prior to the observation of all of the glycans in the sample we know that there is a fixed number total n present in the sample, but until we observe the sample we do not know the precise number. Likewise, the ratios of each species of glycan are unknown in the initial sample until measured. One way to quantify the nature of this uncertainty is through the use of the multinomial distribution.

Modeling the data using the multinomial distribution allows for computation of the likelihood of making specific observations under a specific allocation of the parameters. This likelihood determines the probabilities of making observations of a specific species of glycan. There has been a significant history of analytical manipulations of the multinomial distribution that allow for the application of the model to answer different questions. One such useful application of the model is the development of large sample confidence intervals [Quesenberry and Hurst, 1964][Goodman, 1965]. The confidence intervals themselves are derived as the solutions to the following equation:

$$(\hat{p}_i - \pi_i)^2 = A \cdot \frac{\pi_i(1-\pi_i)}{N}$$

The resultant form of the confidence intervals can be determined by applying Equation 3.1 to the data for the desired proportion interval. A $1-\alpha$ large sample confidence interval

for the proportions of a multinomial distribution is given by

$$\left(\frac{\chi^2 + 2n_i - \sqrt{\chi^2(\chi^2 + 4n_i(N - n_i)/N)}}{2(N + \chi^2)}, \frac{\chi^2 + 2n_i + \sqrt{\chi^2(\chi^2 + 4n_i(N - n_i)/N)}}{2(N + \chi^2)} \right) \tag{3.1}$$

Where i goes from 1 to k with k as the number of categories, n_i is the number of observations of the i^{th} kind, N is the sum total of the n_i values, and χ^2 represents the $\frac{1-\alpha}{2}$ quantile of the χ^2 distribution with degrees of freedom $k - 1$. The modeling approach that is used to describe the glycan experiments uses a multinomial likelihood based approach similar to this method. The glycosylation data is captured by inspecting the areas under peaks on the capillary electrophoresis curves that are associated with three different kinds of IgG samples:

1. Fab - The Fab sub-domain of the IgG molecule, this is the portion of the antibody that binds to pathogenic invaders such as the bacillus *Mycobacteria* tuberculosis.

2. Fc - The Fc sub-domain of the IgG molecule which is separated from the Fab sub-domain in this sample. The Fc subdomain is the portion of the IgG molecule which binds to the lymphocytes.

3. whole - The entire IgG molecule is inspected in this sample, although the sample from which these IgG are isolated is distinct from the samples where the Fc and Fab sub-domains are isolated.

The quantification procedure for the glycan species attached to IgG molecules in terms of the separation of the initial samples as displayed in Figure 3.2, illustrates that the whole glycan species quantification occurs in a different subsample than the Fc and Fab quantification. The different glycans quantified in each subdomain and across the entire IgG molecule in subsamples from the tuberculosis patients are shown in Figure 3.3. The glycans listed in Figure 3.3 contain 18 identifiable categories of glycans, three of which are not uniquely identifiable. In future work, an assessment of the properties of datasets of this type, and potential separation techniques for these categories will be explored, but in this work, these indistinct categories are treated as separate entities from the glycan combinations that they represent.

Initially the dataset reported only the relative concentrations of the glycan data. That is, they contained the respective area under a given peak in the CE curve scaled by the total of all the peaks. The data in this form was not amenable to analysis using a parametric model such as the multinomial, which is a key part of the semi-parametric modeling and prediction framework which we suggest for data of this kind. As such the raw data was requested and provided. The raw data included the actual tallied intensities of each of the peaks on the CE curve.

With the raw data in hand, the ability to produce a parametric likelihood model with this portion of the data became possible. The technique and approach described here in the modeling of this dataset is general and can be applied in the case of any data consisting of relative abundances, modeled as multinomial, and other numerical variables without known distributions. These nonparametric data are most appropriately considered in terms of estimators of their non-parametric densities, such as kernel density estimators which are used in this case.

In the case of the 2021 tuberculosis dataset there were a total of three separate compositions, but there need not be a fixed number of compositions to apply the techniques discussed here. In fact, a previous tuberculosis dataset looked not only at the composition of glycans on IgG molecules in general, as well as their sub-domains, but also on two other specific kinds of IgG which had been isolated, and formed a fourth and fifth composition in the data [Lu et al., 2020].

3.5.1. Model Specification

In this section the theory and specification of the semi-parametric class-likelihood prediction modeling procedure is discussed. First the parametric portion of the model, which treats compositions among the dataset as observations from multinomial distributions, is described. The log-likelihood equations for the model are derived and presented. Following this the modeling procedure for the remainder of the variables (the non-parametric) part of

the likelihood is introduced. The combination of the two parts (the parametric and non-parametric) is discussed, and a general algorithm for training and prediction is provided. Finally the results of applying the semi-parametric modeling procedure to the 2021 tuberculosis dataset are shown, and the confusion matrix of the model for predicting classes of tuberculosis patients is discussed.

Suppose there are a total of K compositions of interest, the term 'composition', here, refers to a collection of variables $(n_1, n_2, \ldots, n_J)$ such that the sum-total of the variables yields a value representative of the magnitude of all observations [Filzmoser et al., 2018a].

For simplicity of argument, consider a bag of multi-colored marbles [Walley, 1996]. The total number of marbles is N. Suppose further that there are J unique colors represented among those contained in the bag, and that for $j \in \{1, 2, \ldots, J\}$, the value n_j represents the number of marbles in the bag of that specific color, such that $N = \sum_{j=1}^{J} n_j$. In the same way, the different subdomains considered (Fc, Fab) and the full molecule IgG may be considered as different types of bags, each of which contain the same kinds of glycans but in different expected ratios. In the case of the 2021 tuberculosis dataset, $K = 3$, and the total number of different glycans within each of the compositions, $J_1 = J_2 = J_3 = 21$. Suppose that there are a total of p distinct classes into which data may be placed. For example, there are three different designations of tuberculosis status: active, latent, and negative, and there are two designations of diabetes status: diabetic and non-diabetic. All of the negative tuberculosis patients were non-diabetic, which makes for five distinctions when considering the combination of the comorbidities. In general we can denote the total p classes using the variable designations $c_1, c_2, \ldots c_p$, and the random variable C for representing the unknown class of a sample.

The multinomial distribution function has a vector parameter that indicates the expected proportion of each kind of outcome of size J_k. The number of possible outcomes for composition k are expressed as J_k. In the case of the IgG glycan compositions, there are 21 distinct peaks that are examined. Each of these may be thought of as a distinct outcome. If of all the glycans in a particular composition, one is chosen at random, the probability it is

of a certain species for all of the species possible is the J_k-vector parameter of the associated composition's multinomial distribution.

For each of the J_k different possible outcomes, let the associated proportion parameters for a particular class j within the k^{th} composition be denoted by $\pi_{j(k)}(C)$. That is, for outcome j of composition k for class C. The complete vector of all parameters for the multinomial distribution can then be represented as the vector $\boldsymbol{\pi_{(k)}(C)} = [\pi_{1(k)}(C), \pi_{2(k)}(C), \ldots, \pi_{J_k(k)}(C)]^T$.

The multinomial distribution is based on observations of count data distributed into the J_k categories. The vector random variable sequence $\boldsymbol{X_i}$ represents a random sample of observations of the composition. For a single observation, $\boldsymbol{X}$, the multivariate probability mass function for a specific outcome class C might be expressed as in Equation 3.2.

$$P(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\pi_{(k)}(C)}) = \frac{\sum_{j=1}^{J_k} x_j}{\prod_{j=1}^{J_k} x_j!} \prod_{j=1}^{J_k} \pi_{j(k)}(C)^{x_j} \tag{3.2}$$

Clearly the only part of Equation 3.2 that depends on the class is the vector parameter $\boldsymbol{\pi_{(k)}(C)}$. The likelihood of the multinomial distribution can be considered as a function of the vector parameter, where the data are fixed values. By considering Equation 3.2 the likelihood function of a series of N observations as $\boldsymbol{X_i}$ can be represented as in Equation 3.3.

$$\begin{aligned} L(\boldsymbol{\pi_{(k)}(C)})|\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_N}) = & \\ & \prod_{i=1}^{N} P(\boldsymbol{X_i} = \boldsymbol{x_i}; \boldsymbol{\pi_{(k)}(C)}) = \\ & \prod_{i=1}^{N} \left(\frac{\sum_{j=1}^{J_k} x_{ij}}{\prod_{j=1}^{J_k} x_{ij}!} \prod_{j=1}^{J_k} \pi_{j(k)}(C)^{x_{ij}} \right) \end{aligned} \tag{3.3}$$

To determine the maximum likelihood estimators of the parameters within each class, the log of Equation 3.3 is easier to work with as it allows for smaller values in the resultant calculations. The log is also a monotonic transform, meaning that the ordering is preserved, and hence optimizing the log of a function results in the optimizing the function as well. The maximum likelihood estimators of the proportion parameters within class C when given a set of samples from a composition's multinomial distribution can be derived from the

log-likelihood function which is given in Equation 3.4.

$$\ell(\boldsymbol{\pi_{(k)}(C)}|\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_N}) = \sum_{i=1}^{N} \left(\log \left(\sum_{j=1}^{J_k} x_{ij} \right) - \sum_{j=1}^{J_k} \log (x_{ij}!) + \sum_{j=1}^{J_k} x_{ij} \log \left(\pi_{j(k)}(C) \right) \right) \tag{3.4}$$

Determining the optimal parameter from the log-likelihood in Equation 3.4, by taking the derivative and solving for the root, while simultaneously imposing the linear constraint that $\sum_{j=1}^{J_k} \pi_{j(k)}(C) = 1$ indicates that maximum likelihood estimators (MLEs) within a given class C may be represented as in Equation 3.5.

$$\widehat{\pi_{j(k)}(C)} = \frac{\sum_{i=1}^{N} x_{ij}}{\sum_{i=1}^{N} \sum_{l=1}^{J_k} x_{il}} \tag{3.5}$$

The MLEs within each of the p total classes are determined from Equation 3.5 applied within each class as
$\widehat{\boldsymbol{\pi_{(k)}(c_1)}}, \widehat{\boldsymbol{\pi_{(k)}(c_2)}}, \ldots, \widehat{\boldsymbol{\pi_{(k)}(c_p)}}$. The determination of these parameter vectors, one for each of the p classes, constitutes the 'learning' or model estimation portion of the prediction method for the parametric part of the compositions.

The prediction consists of computing the log likelihood in Equation 3.4 for each of the estimated MLEs, and determining the highest value. That is for new compositional data $\tilde{\boldsymbol{x}}$, compute $P(\tilde{c} = c_1) = P(\boldsymbol{X} = \tilde{\boldsymbol{x}}; \boldsymbol{\pi_{(k)}(c_1)})$, $P(\tilde{c} = c_2) = P(\boldsymbol{X} = \tilde{\boldsymbol{x}}; \boldsymbol{\pi_{(k)}(c_2)})$, $\ldots$, $P(\tilde{c} = c_p) = P(\boldsymbol{X} = \tilde{\boldsymbol{x}}; \boldsymbol{\pi_{(k)}(c_p)})$, and determine the predicted class $\tilde{c}$ as c such that $P(\tilde{c} = c) = \max \left(P(\tilde{c} = c_t) \quad \forall \quad t \in 1, 2, \ldots, p \right)$.

When there are multiple compositions in the data, as is the case with the 2021 tuberculosis dataset, (recall $K = 3$ in this case, Fab, Fc, and whole), the logs of these probabilities are computed for each composition separately, and added together prior to prediction. That is, for new multi-compositional data $\tilde{\boldsymbol{x}} = [\widetilde{\boldsymbol{x_{(1)}}}, \widetilde{\boldsymbol{x_{(2)}}}, \ldots, \widetilde{\boldsymbol{x_{(K)}}}]$ the class log-probabilities for each of the p classes are computed as in Equation 3.6.

$$\begin{aligned}\log(P_{\text{para}}(\tilde{c} = c_t)) \propto &\log(P(X = \widetilde{\boldsymbol{x_{(1)}}}; \boldsymbol{\pi_{(1)}(c_t)}))+ \\ &\log(P(X = \widetilde{\boldsymbol{x_{(2)}}}; \boldsymbol{\pi_{(2)}(c_t)})) + \cdots + \\ &\log(P(X = \widetilde{\boldsymbol{x_{(K)}}}; \boldsymbol{\pi_{(K)}(c_t)}))\end{aligned} \tag{3.6}$$

The most likely class is then determined by selecting the class associated with the maximum across all t, that is $\forall t \in \{1, 2, \ldots, p\}$.

One of the key assumptions of the modeling approach is that the compositions are actually independent from one another; hence, the probabilities might be directly multiplied without considering the joint distribution. In future work, more consideration of the effect of dependence among the compositions will be investigated. For the tuberculosis data, the independent composition assumption for this model is likely a good assumption, as the samples from which the compositions are computed are actually unique subsamples.

The 2021 tuberculosis dataset contains an additional 19 variables, listed in Table B.1. The probability distributions behind these variables are unlikely to be a standard parametric form, such as a Normal distribution. In lieu of imposing an underlying parametric distribution on the population from which the data are drawn, a nonparametric method of estimating the densities for the remaining variables may be used to determine the most likely class of the data.

The density for the non-compositional part of the data can be determined in a variety of ways. One such way is to smooth the histogram with a particular kernel. This approach is known as kernel density estimation, and the resulting smooth function is called a kernel density estimate (KDE) of the distribution for a particular variable [Wasserman, 2006]. The general form of the kernel density estimator is given in Equation 3.7.

$$\hat{f}_n(z) = \frac{1}{q} \sum_{i=1}^{q} \frac{1}{h} R\left(\frac{z - z_i}{h}\right) \tag{3.7}$$

Observed outcomes are denoted z_i, and the total number of distinct outcomes is given by q. The kernel at location z_i is given by $R(\cdot)$, and the bandwidth, which determines the

interval length along the domain of possible outcomes to consider is h. Several different KDEs (one for each of the p total classes) must be constructed for each variable. Determinations of the class-likelihoods for each new observation z^* can be determined by choosing the value along the x-axis at which the density estimate is maximized.

The density estimate is determined for all variables in each class separately. This model does not consider the non-parametric joint distributions of the data within each class when making classifications. It is left to future work to perform estimation of the joint distributions of the non-parametric data in the model.

Let $\hat{f}_{nl(c)}(z)$ indicate the KDE of the l^{th} non-parametric variable within class c. The 'training' or model-building phase of the non-parametric portion consists of determining the KDE $\hat{f}_{nl(c)}(z)$ within each of the separate p classes, for each of the $l \in \{1, 2, \ldots, L\}$ total variables.

Once these kernel density estimators have been determined, and new data $\tilde{\boldsymbol{x}} = [\widetilde{x_1}.\widetilde{x_2}, \ldots, \widetilde{x_L}]^T$ is available, the class may be predicted by viewing the non-parametric densities within each class as likelihoods. This prediction is independent from that which uses the compositional data, by considering class log-likelihoods as described in Equation 3.8.

$$\log\left(P_{\text{nonpara}}(\tilde{c} = c)\right) \propto \sum_{l=1}^{L} \log\left(\hat{f}_{nl(c)}(\widetilde{x_l})\right) \tag{3.8}$$

The predicted class for the data can then be found by determining the c_t for $t \in \{1, 2, \ldots, p\}$ such that $\log(P_{\text{nonpara}}(\tilde{c} = c))$ is at its maximum.

The parametric and nonparametric prediction can be combined into a unified semiparametric class prediction. However, the magnitudes of the parametric and the non-parametric log-likelihoods may be very different depending on the nature of the data. As such, it becomes necessary to combine the data in a specific way so that one part of the model does not overshadow the other. One way to combine the information from the parametric and non-parametric procedures is to reweight the log-likelihoods such that the contribution to the overall prediction is approximately equal from both portions.

The class log-likelihoods are scaled by the sum of class log-likelihoods in the parametric and non-parametric data separately. Due to the fact that these are always negative, when scaled by their total the maximum becomes the minimum. Therefore, when taking the scaled log-likelihoods from the parametric and the non-parametric portions and adding them together, the classification is made by finding the class which minimizes the sum of the two scaled log-likelihoods. The scores for class c from the parametric and nonparametric parts are represented as $S^*_{\text{para}}(c)$ and $S^*_{\text{nonpara}}(c)$, and the combined score $K^*(c)$, the predicted class c is represented by $\hat{c}$. The scaling, combination, and prediction procedure is described in Equation 3.9

$$\begin{aligned} S^*_{\text{para}}(c) &= \frac{\log(P_{\text{para}}(\tilde{c}=c))}{\sum_{t=1}^{p}\log(P_{\text{para}}(\tilde{c}=c_t))} \\ K^*_{\text{nonpara}}(c) &= \frac{\log(P_{\text{nonpara}}(\tilde{c}=c))}{\sum_{t=1}^{p}\log(P_{\text{nonpara}}(\tilde{c}=c_t))} \\ S^*(c) &= S^*_{\text{para}}(c) + S^*_{\text{nonpara}}(c) \\ \hat{c} &= \{c | S^*(c) = \min\left(S^*(c_t) \ \ \forall \ \ t \in \{1,2,\ldots,p\}\right)\} \end{aligned} \tag{3.9}$$

The complete semi-parametric prediction procedure consists of getting the scores for each of the classes in the parametric and non-parametric models, scaling within each, combining the scores, and determining the minimum across all the classes for the prediction. The training procedures consists of obtaining a MLE for the vector-proportion parameters for each class separately for each of the composition variables and KDE of the nonparametric densities of each class separately. These two components together form a scoring function for each class. The class which produces the minimum score is the most likely class, and serves as the prediction of the semi-parametric model.

The kernel used in KDE of the nonparametric distributions is the Gaussian distribution, but any sufficiently smooth kernel would serve.

3.5.2. Data Study: 2021 Tuberculosis Data - Semiparametric

In the tuberculosis dataset, there are three compositions. Each involves the same 21 glycan species shown in Figure 3.3. In addition to these 63 (21×3), there are 19 numerical variables associated with each patient (described in Table B.1). The three compositions are modeled using the parametric approach. The multinomial likelihood is used to determine the most likely class. The 19 additional variables are modeled using the non-parametric approach, KDEs are computed for each within each class. In this dataset, there are five outcome classes:

1. Active tuberculosis with diabetes (18 patients),
2. Active tuberculosis without diabetes (25),
3. latent tubereculosis with diabetes (7),
4. latent tuberculosis without diabetes (18), and
5. negative for tuberculosis and diabetes (15).

For analyzing the 2021 dataset in addition to other datasets of a similar nature, a general code-base was developed in R that consists of utility functions for the training and prediction procedures of the semi-parametric approach.

For the 83 patients in the full 2021 tuberculosis dataset, the multinomial likelihood functions were trained. The computation took place locally on an x64 architecture Intel Core i7-8550U CPU (1.8 GHz) with 16 GB of RAM, running a 64-bit operating system (Microsoft Windows 10 Enterprise), and took approximately 101 ms. The kernel density estimation of the nonparametric densities for the 19 variables within each of the five classes took around 29 ms with the same system. Unless otherwise indicated, all computations were programmed and evaluated using the R programming language, version 4.1.0. [R Core Team, 2021]

Prediction takes longer than training for this model, as the score functions provided in the training procedures can be computationally expensive. Data repackaging for prediction is also necessary, and accounted for in the prediction times reported here. For the entire dataset, the non-parametric prediction routine takes 53.372 seconds to compute scores for

each of the classes for 83 patients, while the parametric prediction routine takes only 28.29 seconds.

In a first pass approach, the entire dataset is used to train the model, and the model is then assessed on the entire dataset to determine a base-line effectiveness of the model at utilizing information from the data.

Table 3.1: Confusion Matrix for Semi-Parametric Model (All Data)

	True Class	Predicted Class				
		ATB/DM	*ATB/ND*	*LTB/DM*	*LTB/ND*	*Neg/ND*
Nonparametric Score Only	*ATB/DM*	15	1	0	2	0
	ATB/ND	2	17	2	3	1
	LTB/DM	0	0	6	0	1
	LTB/ND	1	0	1	13	3
	Neg/ND	0	0	0	1	14
Parametric (Multinomial) Score Only	*ATB/DM*	16	1	1	0	0
	ATB/ND	10	7	1	5	2
	LTB/DM	3	0	1	3	0
	LTB/ND	0	5	1	11	1
	Neg/ND	1	2	2	2	8
Combined Semiparametric Score	*ATB/DM*	16	1	0	0	1
	ATB/ND	8	12	1	3	0
	LTB/DM	0	0	4	3	0
	LTB/ND	0	0	1	14	2
	Neg/ND	0	1	0	1	12

Table 3.1 displays the confusion matrix when the semi-parametric model is trained using all of the data, and then used to predict the class of each of the observations. Note that when there is data missing, there is no contribution to the score for the respective model. The 83 patients in the 2021 tuberculosis dataset have both their tuberculosis and diabetes disease classes determined in the first prediction model. Due to the fact that there are five possible classifications, if the class were assigned by random guessing, then the expected accuracy would be around one fifth or 20% if the set were balanced. The accuracy of the semi-parametric combination, parametric, and non-parametric models are 72.5 %, 51.2 %, and 78.3 % respectively.

This analysis is trained using the initial classifications of the data. If this information was stored in the model directly during the training the classification rate would be 100 %. The semi-parametric model seeks to build a model which includes the maximum amount

Table 3.2: Confusion Matrix for Semi-Parametric Model (Leave One Out CV)

	True Class	Predicted Class				
		ATB/DM	*ATB/ND*	*LTB/DM*	*LTB/ND*	*Neg/ND*
Nonparametric Score Only	*ATB/DM*	7	9	0	2	0
	ATB/ND	7	12	2	3	1
	LTB/DM	0	3	0	2	2
	LTB/ND	1	4	1	9	3
	Neg/ND	1	0	3	5	6
Parametric (Multinomial) Score Only	*ATB/DM*	15	2	1	0	0
	ATB/ND	10	4	1	5	5
	LTB/DM	3	0	0	3	1
	LTB/ND	0	5	2	10	1
	Neg/ND	1	5	2	2	5
Combined Semiparametric Score	*ATB/DM*	14	3	0	0	1
	ATB/ND	11	8	1	3	1
	LTB/DM	2	2	0	3	2
	LTB/ND	0	3	3	9	2
	Neg/ND	1	2	2	4	5

of information without over-constraining the model. To assess the ability of the model to predict the disease class, an additional validation procedure is applied. Leave-one-out cross validation, a jack-knife approach where statistics are computed from the entire dataset with the exception of a single hold out, is applied to the dataset in the same manner, and the results are displayed in Table 3.2. The accuracy using the leave-one-out cross validation approach is 41%, for parametric and nonparametric scores and 43% for the combined score. This indicates a potential gain in predictive capability when the scores are combined.

In addition to predictions that are made on the five class split of the dataset, predictions of only the tuberculosis class, and only the diabetes class were also made using the same procedure. The results of these classifications using the entire dataset are given in Tables B.3 and B.5.

The accuracy when considering only the tuberculosis status in training and in prediction for the parametric, nonparametric, and combined semi-parametric scores using the entire dataset for the training were 65%, 80%, and 84% respectively. In this case, random guessing for the prediction of the tuberculosis status would result in an expected accuracy of 33%, indicating that a considerable amount of valuable information can be learned and utilized by the semi-parametric modeling and scoring approach for prediction. When the semi-parametric approach is validated using a leave-one-out approach the resulting confusion

matrix is as in Table B.4.

The accuracy computed from the leave-one-out cross validation of the semi-parametric model for classification of the tuberculosis status only were 57% (parametric), 58% (non-parametric), and 65% (combined). In this case, there were three cases where the prediction scores were equivalent, and these are removed from the confusion matrix, hence there are only 80 cases for analysis in the bottom portion of Table B.4.

In addition to classification of the tuberculosis group status, the diabetes status of the patients (either diabetes present or not) is also classified using the semi-parametric modeling approach, when the model is trained using all of the data, the results are as given in Table B.5. The accuracy of the parametric, nonparametric, and semi-parametric models in classifying diabetes status when trained using the entire dataset were 72%, 84%, and 79% respectively. The results of the leave-one-out cross validation of the model for diabetes status are in Table B.6.

The accuracy for the nonparametric, parametric, and semi-parametric models for prediction of diabetes status among the leave-one-out cross validation were 62%, 67%, and 68% respectively. For a visual representation of the differences among the five-classes in the non-parametric data, the kernel density estimators for each of the 19 variables are displayed in Figure B.1.

From the prediction results, the effectiveness of the pieces of the semiparametric model in the 2021 tuberculosis dataset at determining the disease class for tuberculosis and diabetes can be seen. The nonparametric approach tended to provide better information about the disease outcome classes, which indicates that the nonparametrically modeled variables in this case provided a stronger partitioning of the data into these groups than the parametrically modeled ones did. This is likely not to be a generally true statement for all such data sets. It is also clear that in several cases the combination of the class-likelihoods from the parametric and the nonparametric models tends to increase the overall accuracy of the model in prediction, indicating utility in combining the two scores for prediction.

3.6. Glycan rank probability method

Another nonparametric model is suggested here for investigating the composition data. This model is based on the probabilities that the relative abundances of glycans will have a specific ordering (for all pairs) within a given class. In the "bag of marbles" scenario, suppose that two individuals (A and B) have multiple bags of marbles. Given a randomly selected bag of marbles, the colors of each marble are quantified and compared to show there are more reds (n_r) than greens (n_g), and more greens than blues (n_b) in the bag. Now suppose A and B both provide three bags of marbles, and during the quantification process of each of these, it is found that the proportions from the unknown bag ($n_r > n_g$ & $n_g > n_b$) are much more consistent with the patterns exhibited by marble-bags belonging to B. A determination therefore is made to designate the predicted owner of the unknown bag as B.

In the 2021 tuberculosis dataset [Lu et al., 2021], the glycan species pair ranks can be modeled as belonging to a particular disease outcome class. This approach is introduced to attempt to make use of information that might be able to serve as a proxy for the genetic differences of the patients related to glycan species converting enzymes. Biologically, the presence or absence of enzymes that convert one glycan species into another one could account for the variation in the relative rankings of pairs of species. These differences may also account for the variations in the disease responses. If so, probability models based on the glycan ranks within each class may allow for the early differentiation of disease type outcomes.

3.6.1. Link of glycan rank probability to enzyme concentration

The chemical model discussed briefly in this section describes a potential biologically driven model based on the efficiency to create specific enzymes. The Michaelis Menten equations [Michaelis and Menten, 1913, Johnson and Goody, 2011] are a set of differential equations used in biochemistry that relate the concentrations of products and the reactants in enzy-

matic relations. One manner of making determinations on disease subtypes from relative abundances of glycans involves the estimation of enzyme concentrations via observed substrate and resultant compound concentrations. In an enzymatic chemical reaction such as those applied in the conversion of one *species* of glycan to another, an enzyme reacts with a substrate to form an intermediary enzyme-substrate complex. This creates the product, and regenerates the initial enzyme, which is not consumed during the reaction. Symbolically this kind of chemical reaction may be displayed as shown in Equation 3.10.

$$E + S \rightleftharpoons ES \rightarrow E + P \tag{3.10}$$

In the Michaelis Menten equations, the concentrations of the various chemical components are denoted using square brackets, for instance, $[x]$ is the concentration of x in a solution. Furthermore the rate of transition from a solution containing the separate enzyme E and substrate S into a solution containing only the enzyme substrate complex ES is denoted as k_f and the rate of transition back is k_r (f for forward, and r for reverse). The rate of transition from the enzyme substrate complex into the enzyme and the product P is known as the catalytic rate and is denoted by k_{cat}.

The rates involved in Formula 3.10 can be described via a set of related rates equations. These are derived from Cato Guldberg and Peter Waage's law of mass action which was initially conceived of and proposed in the 1860's [Waage and Guldberg, 1864]. In essence, the law of mass action refers to the kinetic action of nearby masses, and specifically refers to the rate of reactions of molecules in a chemical reaction. By maintaining the ratio of the product to the reactant molecules as constant, the rate of reaction can be shown to be directly proportional to the product of the concentrations of the reactants.

The main takeaway from this discussion is that the concentrations of substrate and product (S and P) can be related to the concentration of a catalyzing enzyme. Therefore, one relevant set of observations that may be useful in discriminating between classes that are suspected of having different levels of enzymes available is the observation of the direction of an imbalance in each reasonably related pairing of the glycan species.

3.6.2. Model Specification

In a similar manner to the maximum likelihood classification method discussed in the previous section, the rank probability method also predicts the most likely class. This time, instead of using parametric likelihoods over all of the composition variables simultaneously, this approach places a Bernoulli distribution on each possible pair of transitioning elements that can transition into each other by a single addition or removal of a single simple sugar.

In this model, three internal relationship graphs are built during training and used for prediction/classification. Since there were three compositions where these relative concentrations are reported, three 21×21 element matrices (441-elements, only 116 of which made biological sense) are estimated within each class. These provided point estimates of the probability that one glycan would occur in greater concentration to another within each of the possible outcome categories. In this analysis only the five category (tuberculosis/diabetes) classification was examined. To compute the glycan rank probabilities within each class a simple algorithm was applied.

1. Training Phase: n observations and associated class labels are provided.
    (a) For each of the classes,
        i. For each composition,
            A. For each composition element, compute and save the vector of indicators indicating whether this composition element is more abundant than each of the others.
        ii. Compute the average in each class-composition matrix (to compute and store the proportion estimate per class).
    (b) Return set of rank ordering probability estimates for each class and composition.
2. Prediction Phase:
    (a) For each of the classes,
        i. For each composition,

A. Determine the product of the probability estimates from each class for the observed glycan concentration level (where each of the possible combinations contributes either the rank probability [if A is greater than B] or one minus the rank probability [if B is greater than or equal to A]).

B. Predict most likely class, record this, and return it along with the probability (for combination with other composition variables).

ii. Compute the most likely class after multiplying all within class composition probabilities together.

iii. Return probabilities and predictions.

3.6.3. Model Application Results

The results of applying this algorithm to the IgG glycan data from the 2021 tuberculosis dataset, training with all of the data, and validating on the same data, provides a baseline accuracy for the model. The results of five category classification are in Table 3.3. Using only the whole, Fab, and Fc glycans gave average accuracy of 57.3, 58.2, and 58.6, and the combined accuracy was 65.3%.

Classification when using only the five most variable composition rank orderings in terms of the probabilities within each class is also performed. the results of this five category classification after training with the full dataset are given in Table 3.4.

The accuracy for each of the four different models built on the various components when using only the most highly varying five transitions across the five possible classes were whole - 50%, Fab - 43%, Fc - 48%, and combined 55%.

The results of applying the glycan rank probability model is buffeted by missing data for some glycan species. The model is further plagued by left censoring (where detection intensities fall below the limit of resolution of the machine). In spite of these pitfalls, the glycan rank probability model performed a decent classification into the five groups when data was available. The results of these classifications were comparable to the results of the

Table 3.3: Confusion Matrix for Glycan-Rank probabilities (full data)

	True Class	Predicted Class				
		ATB/DM	*ATB/ND*	*LTB/DM*	*LTB/ND*	*Neg/ND*
whole Glycan Only	*ATB/DM*	14	1	0	0	3
	ATB/ND	4	5	1	3	6
	LTB/DM	1	0	4	0	2
	LTB/ND	0	3	3	8	3
	Neg/ND	1	0	0	0	12
Fab Glycan Only	*ATB/DM*	12	3	0	0	1
	ATB/ND	8	7	0	3	6
	LTB/DM	0	0	5	0	2
	LTB/ND	1	0	4	7	4
	Neg/ND	2	1	1	2	8
Fc Glycan Only	*ATB/DM*	12	1	1	0	2
	ATB/ND	6	6	2	5	4
	LTB/DM	0	0	4	1	1
	LTB/ND	0	0	5	11	1
	Neg/ND	2	1	0	2	1
All Glycan compositions	*ATB/DM*	14	1	0	0	1
	ATB/ND	6	10	0	4	3
	LTB/DM	0	0	5	1	1
	LTB/ND	1	0	1	12	2
	Neg/ND	2	0	1	2	8

semi-parametric model, which indicated that the information carried in the glycan proportions is usable for classification, although the results of application of the glycan probability model were validated using the same data that the proportions were captured from, while the semi-parametric model was validated using leave-one out cross validation.

Table 3.4: Confusion Matrix for Glycan-Rank probabilities (high variance five only)

	True Class	Predicted Class				
		ATB/ND	*ATB/DM*	*LTB/ND*	*LTB/DM*	*Neg/ND*
whole Glycan Only	*ATB/ND*	1	7	9	1	1
	ATB/DM	0	14	1	1	2
	LTB/ND	0	2	10	2	3
	LTB/DM	0	4	1	1	1
	Neg/ND	0	0	1	1	11
Fab Glycan Only	*ATB/ND*	1	11	8	4	0
	ATB/DM	0	11	0	5	0
	LTB/ND	1	1	12	2	0
	LTB/DM	0	0	2	5	0
	Neg/ND	1	2	8	3	0
Fc Glycan Only	*ATB/ND*	9	6	6	1	1
	ATB/DM	3	12	1	1	0
	LTB/ND	4	0	12	1	0
	LTB/DM	2	1	3	1	0
	Neg/ND	3	5	2	0	3
All Glycan compositions	*ATB/ND*	2	6	6	1	3
	ATB/DM	1	12	0	2	1
	LTB/ND	2	0	12	2	0
	LTB/DM	0	1	2	4	0
	Neg/ND	0	1	2	1	8

3.7. Conclusions

In this chapter, several approaches for differentiating disease outcome classes for combined compositional and non-compositional data are suggested. The discussed statistical likelihood procedures are applied to a 2021 tuberculosis dataset [Lu et al., 2021]. The set contained the quantification from CE of glycan species on IgG molecules across three specific subdomain samples for tuberculosis patients. The tuberculosis patients in the dataset were divided into active, latent, and negative tuberculosis subtypes, and their diabetes statuses were also reported. In this dataset, all 15 patients who were negative for tuberculosis were also negative for diabetes; therefore, there were a total of five disease categories for the classification procedures.

First the results of applying a classical method known as PLS-DA to the problem are supplied in the form of visualization and the overall model accuracy when applied with varying numbers of latent variables. The procedure itself performs the classification decently, and appears to show separation in the clustering of the data in visualization. Of the three

methods presented, this approach offers the key ability to visualize the data by reducing the dimension. This procedure also is an approach which has seen varied and widespread usage in a variety of applications.

Secondly a semi-parametric model using the multinomial distribution to describe the glycan species abundance within a sample, and kernel density estimators for the remaining variables was introduced. This approach combines two powerful techniques for estimating the likelihood of a particular observation given a known class outcome for disease. On application of this technique within the tuberculosis dataset it was found that the parametric and nonparametric portions of the data tended to produce different classifications when regarded on their own, which separately were less accurate than when combined. In this work the combination of the parametric and the nonparametric scores is undertaken in a manner which assigns equal weight to the information provided by either portion of the data. It is likely the case that weighting should be applied in a nonbalanced way. As a future direction of study, the application of weights to the scores which will allow for different importance being placed on the parametric and the nonparametric portions will be investigated.

Finally a model that attempts to use the glycan species concentration pair rankings in the dataset as a proxy to the concentrations of special enzymes which may be closely related to the disease process is discussed. The upside to this approach is that it is tied to biological reasoning for differential disease process outcomes. A large downside to the approach is that the 19 non-compositional variables in the dataset are not utilized in the predictive procedure. Indeed, a manner in which to include these observations in the modeling and prediction routines is left to future work in this project.

Overall, the approaches discussed here are useful statistical techniques for differentiating tuberculosis and diabetes subtypes in diseases. The approaches are also more general than the application which is provided here. In some cases it is possible that these approaches might be more powerful during prediction. Another slight limitation of the studies in this section is the data-set size for the 2021 tuberculosis data. Although there are 83 patients

available in the dataset, there are a total of 82 numerical variables of interest, and those are just the ones mentioned here. This very wide dataset inherently has some restrictions in its use; however, it is a great example of a medical dataset such that might be available to a doctor when making a decision on the treatment regimen of a tuberculosis patient.

3.7.1. Future Directions

There have been several future directions mentioned in the body of this chapter. For the PLS-DA procedure several simulation studies can be performed to determine attributes of the approach, as well as its robustness to violations of underlying assumptions of linear relationships. Another direction that is suggested in this work is the application of the PLS approach to impute missing values in the dataset. For example, if the $\boldsymbol{X}$ variable block was considered to be all of the observations for which there is complete data, and the $\boldsymbol{Y}$ block to be the block with some missing observations, then the model can be trained using the observations where all $\boldsymbol{X}$, and $\boldsymbol{Y}$ variables are available, and the missing values of $\boldsymbol{Y}$ may be imputed by their predicted values from the PLS modeling equations.

In the second part of this chapter the semi-parametric model is introduced and discussed. In extending this model, a thorough analysis of the combination of scores from the two portions should be investigated. In other words, simulation studies investigating different methods for combining the information from the two portions in a composition/non-composition dataset will be explored.

Lastly, an approach for utilizing the composition rank probabilities to find the most likely class is presented, for this approach a critical limitation was the inability to directly model the non-compositional variables in a way that corresponds to the compositional approach. As an alternative to the rank probability calculations described, the non-parametric estimation of the density of all pairwise ratios of glycans in the dataset will also be explored in future work. The ability to differentiate patient classes will be compared to the binary probability estimation procedure described in this work.

Chapter 4

Analysis of Epigenetic Signals of Genomic Sequences

This chapter describes a third kind of modern biological data that an investigator may collect: epigenetic signals. Epigenetic cues along genetic sequences contain information that is "on top of" that contained within the genomic sequence itself. Modeling the epigenetic information can be very important in differentiating genetic sequences which are otherwise indistinguishable. There are several kinds of chemical signals which may be present in the genome that indicate that certain actions need to be taken or not taken at a given location. Kinds of epigenetic modifications on the genome include DNA methylation [Bird, 2002] [Jones and Liang, 2009], histone modification [Karlić et al., 2010]
[Henikoff and Shilatifard, 2011], regulatory feedback [Tomikawa et al., 2012]
[Yao et al., 2019], and more [Alberts et al., 2002].

DNA methylation conveys information in the methylation status of particular locations along a genomic sequence. Methylation refers to the chemical addition of a methyl group (R-CH_3), and from a functional perspective indicates the repression of a particular region of a genetic sequence. One recent study has shown that differentially methylated regions (DMRs) can be used to determine information about the age of certain sequences [Bell et al., 2012]. Another study indicated that methylation patterns show pronounced changes in smokers depending on the dose and duration of smoking [McCartney et al., 2018]. Other studies have characterized phenotypic trait differences due to epigenetic modification through methylation rather than transcription and translation [Leenen et al., 2016] [Toyota et al., 1999] [Marsit et al., 2006] [O'Brien et al., 2006] .

4.1. DNA Methylation

The Central Dogma of Molecular Biology states that genes are transcribed into RNA which is translated into protein, and that proteins are responsible for doing the cellular work that creates phenotypes [Alberts et al., 2002] [Hansen et al., 2012]. It has since been proven that additional information about the manner in which the genetic code should be read and manipulated by the organism to best increase their chances of survival and genetic propagation is contained in secondary cues differentiating one form of a base nucleotide from another [Alberts et al., 2002]. For instance the base nucleotide cytosine may occur in a form which is methylated. The methylated version of cytosine indicates to the cell important attributes of the particular region of the genome on which they are found.

These secondary cues are also known as epigenetic factors [Medvedeva et al., 2015], as they are factors important for the encoding and survival of life. However, these cues are not a part of the primary genetic code given in the sequential variation of nucleotides along the DNA. Rather, the epigenetic factor serves as what could be considered a footnote in the overall organization of the genetic code for the organism [Alberts et al., 2002]. The footnote indicates important information about each of the locations which they mark. They may also be seen as landmark flags along the genome that colorfully display the positions of important locations, such as the location of genes that should not be presently translated into RNA for further transcription into protein.

For positions marked with the epigenetic cue of methylation, this is precisely the case. These regions indicate genes that should be silenced in the overall maintenance of the cell. In other words, regions of methylation among the same organism or different organisms along the same or similar regions of their genomes indicate that these genomic regions are handled differently by the translation and transcription apparatus available to the cell, due to the differential methylation patterns on the two unique genomes [Alberts et al., 2002]. Metaphorically, as the series of base nucleotides represent the sentences in a story, read in a specific order to give meaning, these secondary cues may be considered as the punctua-

tion and accents which indicate the relative importance of the sentences. More specifically, base nucleotides or histones nearby might be acetylated (indicating higher import placed on expression of the gene)[Liu et al., 1999], phosphorylated (indicating positions of relevance during replication via mitosis or meiosis) [Blom et al., 1999], methylated (indicating high import on the repression of the gene) [Bird, 2002], or contain various combinations of the different cues in a patterned way indicating other reading-actions [Alberts et al., 2002]. Treating a genomic sample harvested for sequencing with a chemical agent causes changes in the structure of the sample which may be exploited by computational algorithms that allow the investigator to gain a broader general view of the elements that uniquely encode the organism under study.

4.2. Some Previous Statistical Approaches to Methylation

Methylation along a reference provides an indispensable picture of the structure of the genetic material being investiagated. In one sense, methylation profiles display regions where genes are repressed. The utility of the methylation profile (regions along the reference sequence which are found to be methylated with specific proportions of the sample reads) is enhanced by the fact that many of the parts of the genome which are methylated in the parents of the organism under study are propogated to the offspring. These regions may tend to be diverse in a population, specifically in humans due to the genetic cross over of the two progeniters during the procreation of offspring. In this regard, the methylation profile of an organism may be generally used to refine the selective group of potential donors of the material under study. Bisulfite is a chemical agent which may be applied to genetic material prior to analysis which will modify the methylation patterns of the genetic material such that Cytosine which is methylated will remain unchanged, and those which are not methylated will undergo a change such that they are determined to be Thymine by sequencing technologies. Alignment of these reads to a reference sequence and subsequent differential analysis of Cytosine/Thymine mapping allows the user to determine the relative proportion of the unmethylated cytosines to those methylated at a specific CpG or other Cytosine location of

interest in the reference.

4.2.1. CpG point estimates of methylation

A reference sequence generated from the assembly of untreated genetic material harvested allows for a full four-base representation of a genomic sequence of interest, which serves as a background against which to compare genetic reads that *have been treated* in order to determine the secondary cue structure of the material.[Zheng-Bradley et al., 2017] In the case of Bisulfite sequencing, bisulfite treated reads are aligned to a reference sequence by one of several different techniques which allow for mismatches at expected locations (that is allowing certain Cytosines in the reads to be matched with Thymine in the reference and vice-versa) [Wu et al., 2015]. Next the aligned reads are compared to the original four base reference which allows for the determination of methylation status at particular regions along the reference. Specifically, where there are mismatches in the bisulfite treated read indicating a Thymine where there is a Cytosine in the reference methylation is assumed to not have occurred, whereas when there is a perfect matching of the cytosines in the bisulfite treated reads to the reference, methylation is assumed to have occurred. This allows for an estimation of the proportion of donor genetic material which is methylated at specific locations along the reference. [Genereux et al., 2005]

Symbolically, for a reference containing n total CpG or CpH sites (cytosines followed by guanines, or arbitrary bases) indexed by i, let the coverage of the i^{th} CpG/CpH be B_i, a value indicating the number of times that the i^{th} CpG/CpH is 'hit' by reads. For each of the n total locations, of each of the B_i reads that cover that location some subset will indicate a methylated base at that location and others will indicate unmethylated bases. Let U_i be the number of reads that indicate the base is unmethylated, and M_i be the number of reads covering CpG/CpH i that indicate that the base is methylated such that $B_i = U_i + M_i$. Assume that there is a true population proportion of methylation for each CpG/CpH along the reference, such that if a read is selected uniformly from all those possible in the population

of interest that the **true** probability that the read will indicate methylation at CpG/CpH i is given by π_i. Then for any given sample of reads, the maximum likelihood estimate of the true probability of methylation π_i may be computed as $\hat{\pi}_i = M_i/B_i$. These estimates $\hat{\pi}_i$ are called the point-estimates.

Note that n must be an non-negative integer value, as must all of the n total B_i, and the constituent M_i and U_i. Furthermore, it is obvious that each of these represent count values, in the case of n this value is fixed prior to sample analysis, hence it is a standard variable, whereas B_i, M_i and U_i are random variables which are unknowable until they are counted. That said, these random variables *do* depend on the total number of reads that are available in a sample which may be seen as another fixed value, say n_r. The total number of reads that are analyzed in a particular experiment can be controlled, but in practice it is quite difficult to attain uniformly exact coverage at a fixed level across the entire reference. The read length (in terms of bases per read) can also often be controlled, however the control of these two values alone does not guarantee that a specific coverage level will be attained exactly, and hence it is often convenient to represent the coverage using a random variable as we do in this treatment. The coverage is however capped by the number of reads, and this can be expressed in the distribution. Furthermore, the number of reads that express methylation at a particular base and likewise those expressing unmethylated bases at the same loci have induced upper bounds that are interdependent. Specifically, if there are a total of n_r reads, and the desired coverage of a specific sample is set at a proportion $c \in [0, 1]$ then the natural probability distribution of each of the B_i depends on the distribution of the reads across the reference. Let the length of reads in a particular sample be held constant at n_k, and the length of the reference from which the reads are taken be represented by r then each of the reads in the sample (n_r total) begins at location R_j for $j = 1 \ldots n_r$. The distribution of the R_j is another discrete probability distribution which is related to the distribution of coverage for each location. R_j, as the starting position for each read that is contained within a given sample have probability distributions on one to the total length of the reference less n_k. More explicitly, if the length of the reference is $\mathcal{L}$ then there is some probability associated with each of the integer values

between one and $\mathcal{L} - n_k$. Mathematically the distribution may be described in terms of a probability mass function p_R such that $p_R : \{x \in \mathbb{Z} | 1 \leq x \leq \mathcal{L} - n_k\} \mapsto [0, 1]$, where $\sum_{i=1}^{\mathcal{L}-n_k} p_R(R_j = i) = 1 \quad \forall j = 1, 2, \ldots n_r$.

Usually this distribution cannot be precisely controlled by the technology used to capture the reads from a particular sample. Generally, it is assumed that this distribution will neatly be concordant with a discrete uniform distribution, however, due to many simultaneously unmeasured effects of the configuration of the genetic material, the sampling technology used, as well as relative qualities and filters it is unlikely that the distribution of reads will exactly follow the uniform distribution in practice. In other words, the discrete uniform distribution is a *convenient* model for the distribution p_R however, in practice it is not exactly correct rather, minor perturbations here and there in terms of the probabilities that a read will start at a specific location are potentially more correct yet more difficult to model and require more computational space requirements.

Of course, the distribution of sampled reads in practice will never exactly meet the coverage that is desired for a specific experiment, if the desired coverage is say 100 times, then there will be some portions of the reference which are covered slightly fewer, and others which are covered slightly more.

The BSmooth software by Kasper Hansen, Ben Langmead et al [Hansen et al., 2012] performs a local likelihood smoothing [Loader, 1999] of the point estimates of methylation for a given bisulfite sequencing experiment [Frommer et al., 1992]. Conceptually the smoothing procedure follows a traditional local-likelihood procedure.

The code itself is contained in the Bioconductor [Morgan, 2021] software suite for R [R Core Team, 2021]. There was a study in 2015 which sought to compare the `BSmooth` procedure to far simpler techniques when it comes to estimating the average methylation rate over a given area. It was effectively shown in that work that the BSmooth procedure did not really provide a vast improvement over the far more simplistic procedures of moving averages in the case of estimating the average methylation along a genome. [Wu et al., 2015] Another software provides a self-contained analysis suite for the methylation profile of ge-

nomic signals. DMAP (the software - Differential methylation Pattern Analysis) [Stockwell et al., 2014] is marketed as especially useful in the analysis of Reduced Representation Bisulfite Sequences (RRBS) and Whole Genome Bisulfite Sequencing (WGBS) experiments.

4.3. Potential Methylation Locations along A Reference (GRCh 38)

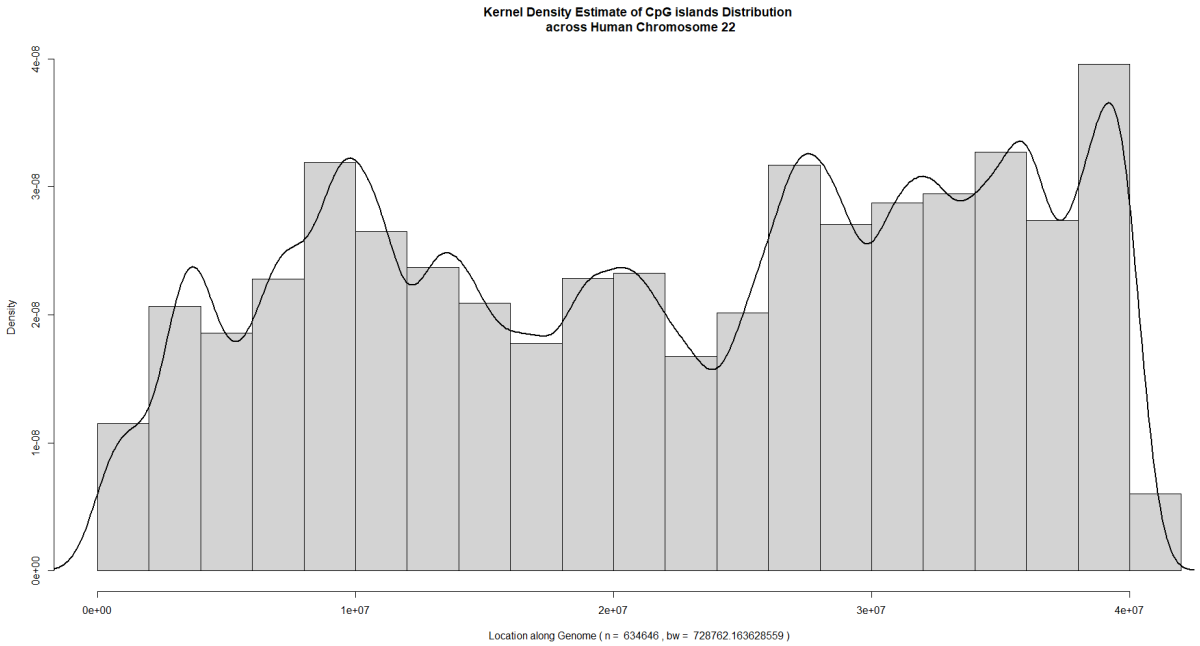


Figure 4.1: KDE of CpG Island Distribution on Human Chromosome 22 (GRCh38)

The human reference genome is a representative structure arrived at by consensus of a large sample of humans, in this study the GRCh38 version is used. First a map of all locations where a CpG island exists is constructed. A simple algorithm which checks all subsequent pairs of locations as "c" and "g" respectively determines the locations, and the CpG density of Chromosome 22 may be inspected. The techniques investigated here will primarily have to do with the techniques of smoothing; therefore, a kernel density estimator with a Gaussian kernel using a bandwidth of 728,762 locations is displayed along with a histogram of the number of CpGs along the reference genome of Chromosome 22 in Figure

4.1. It is apparent that the relative abundance of CpG islands, and therefore potentially methylated regions, varies over chromosome 22. There is a tapering off in abundance at both ends of the sequence, as well as a small lull in density near the center. Examining these CpG density patterns prior to sequencing may allow for slightly more targeted studies which seek to investigate differentially methylated regions only along the most densely CpG dominant regions.

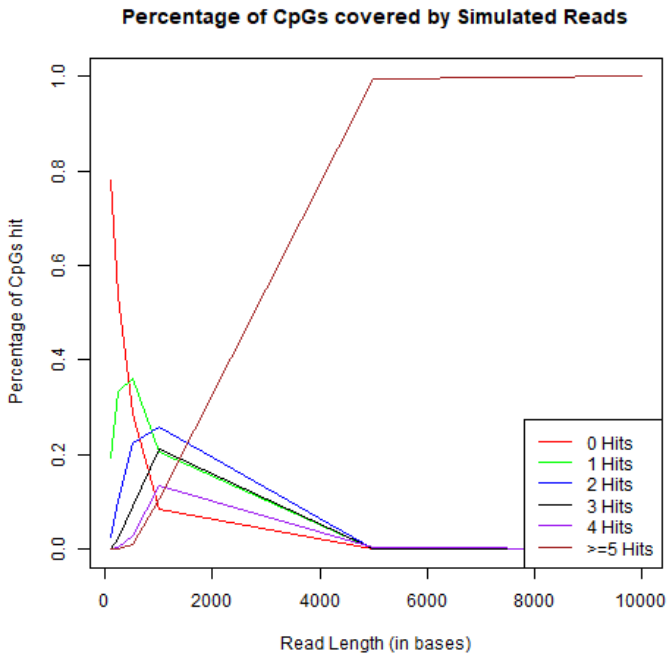


Figure 4.2: Number of reads covering CpGs in human Chromosome 22

Following the determination of the CpG locations along Chromosome 22, a random sampling procedure is applied with read lengths of varying sizes (100,250,500,1000,5000, and 10000 bases) In order to determine the relative coverage of 100,000 reads over Chromosome 22. The Chromosome itself consists of long tails of non-identifiable bases represented by 'N' in the FASTA format, which is not unlike other chromosome structures. For each of the read lengths and for all simulated reads (100,000) the number of times a CpG was 'hit' by a read are captured. For human chromosome 22, there are a total of 634,646 CpG locations. The total number of identifiable bases in chromosome 22 is 40,298,466, indicating that roughly

1.5% of chromosome 22 may exhibit differential patterns of methylation. As the read size increases, if the reads are sampled uniformly at random from the chromosome, the total number of hits as well as the proportion of reads hit more than once increases as displayed in Figure 4.2.

Figure 4.2 shows that as the length of the reads increases the number of CpGs covered by a larger a number of reads increases. While this seems somewhat obvious, it is illustrative of an interesting side-effect of smoothing out the methylation profile. However, perhaps more illustrative of one of the problems investigated here is displayed in Figure 4.3, which shows the trend of variance in the number of hits per CpG. Moving forward, it should be straightforward to produce a purely analytical proof of the relationship between CpG Hit variance and read length for this uniform sampling paradigm.

It makes significantly more sense to look at a smaller region of the genome that will be targeted for sequencing rather than the entire chromosome in a smaller scale simulation study of this kind. However, proper consideration should be given to the particular region of interest for a specific experiment, in the case of this simulation study regions containing around 5-10 genes will be specifically targeted. In addition to being more accurate in terms of a real-life experiment, partitioning at this level will allow for finer granularity in the resultant.

The simulation suggests, per every single base extension in read length, that on average the variance in the number of hits on each of the 634,646 CpGs increases by 0.002502 (SE 7.869 $\times 10^{-6}$). The implication of the linear trend here is that as the length of reads (or effectively the number of reads) is increased the greater the chance for more dramatic changes in the coverage over the length of the chromosome. The simulation next generates a synthetic ground truth point methylation profile for chromosome 22 by randomly sampling a uniform distribution from zero to one 634,646 times. Next, for each of the CpGs, depending on the number of reads that hit them, and the synthetic ground truth methylation point proportion for that particular CpG, a simulated count of reads indicating methylation at that position are generated, and point estimates of the methylation proportion at each position

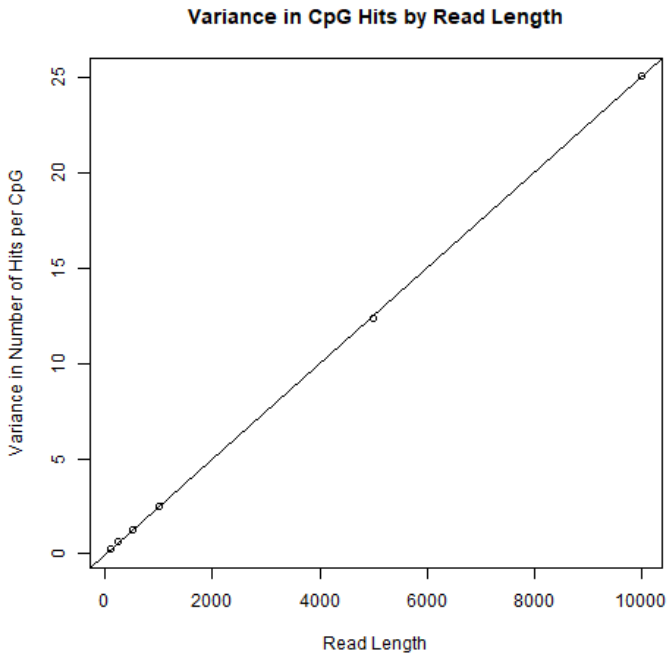


Figure 4.3: Variance in number of CpG "Hits" per Read Length

are computed. The average error of these estimates for a single run of the simulation is displayed for varying read sizes in Figure 4.4. Moving forward, a method of speeding up the simulation will allow multiple simulations and a display of the simulated distribution of this total error.

Following the simulation and estimation of methylation point proportions for the chromosome based on the number of hits and a synthetic ground truth of methylation, the methylation point estimates are smoothed using a constant window size. This is a key difference from the BSmooth procedure which uses an adaptive window size. The average error is computed for each smoothing based on the ground truth, for the simulated 10,000 base length read data. The results are displayed in Figure 4.5.

When applying the constant window size smoothing approach, we appear to be incurring a bias in the methylation point estimates while decreasing the variance of the estimates over time. This is manifest in the increased error as the window size increases. In the current simulation script the methylation proportions in the ground truth synthetic data are

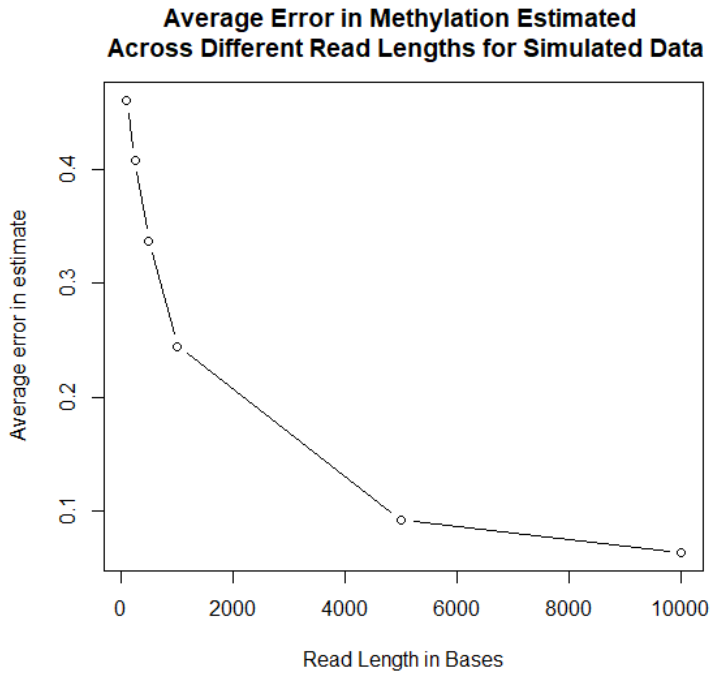


Figure 4.4: Average Error in Methylation Point Estimates over Varying Read Lengths from simulated ground truth

independent and uniformly generated from 0 to 1. In reality there are likely correlations between these proportions, which will be more thoroughly investigated by placing patterned correlations in the synthetic ground truth point methylation estimates in the simulation.

4.4. Reduction of Multi-Mapped Reads

In the analysis of sequencing data, when a file of reads are to be mapped to a reference, it is possible that a read will align to a single position. Frequently, the read will align to a few positions, especially if the reference and read are repetitive. An application of smoothing procedures is the ability to differentiate each position's estimate. This is helpful in assigning different scores to the positions later, when attempting to determine the most likely position from among a set of possible positions later.

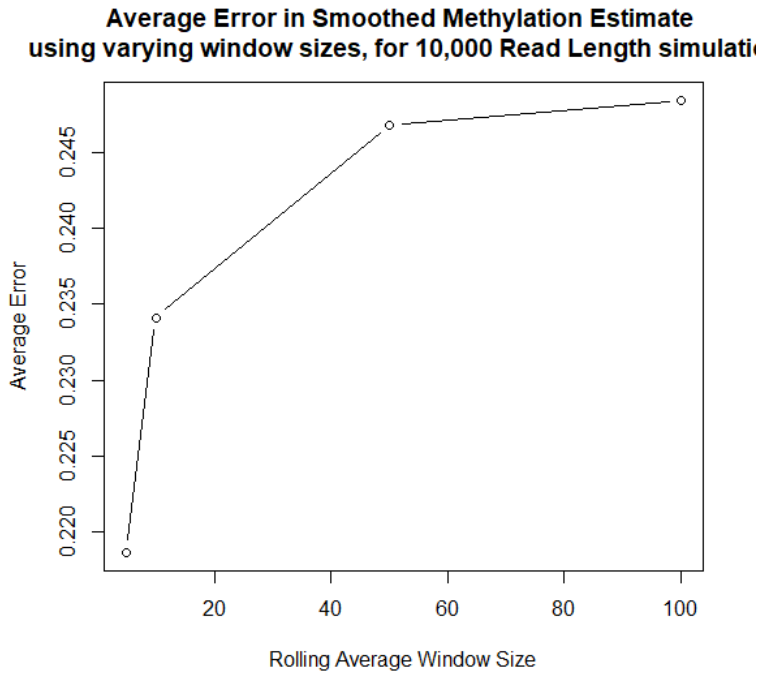


Figure 4.5: Average Error for windowed average smoothing approach on Simulated reads with synthetic ground methylation data

If only the point estimate were used in this case, it is likely that when the coverage is relatively uniform across the genetic sequence there would be many proportions that were identical. The possible locations within the set for a multimapped read might indicate the same probability that the read came from this location.

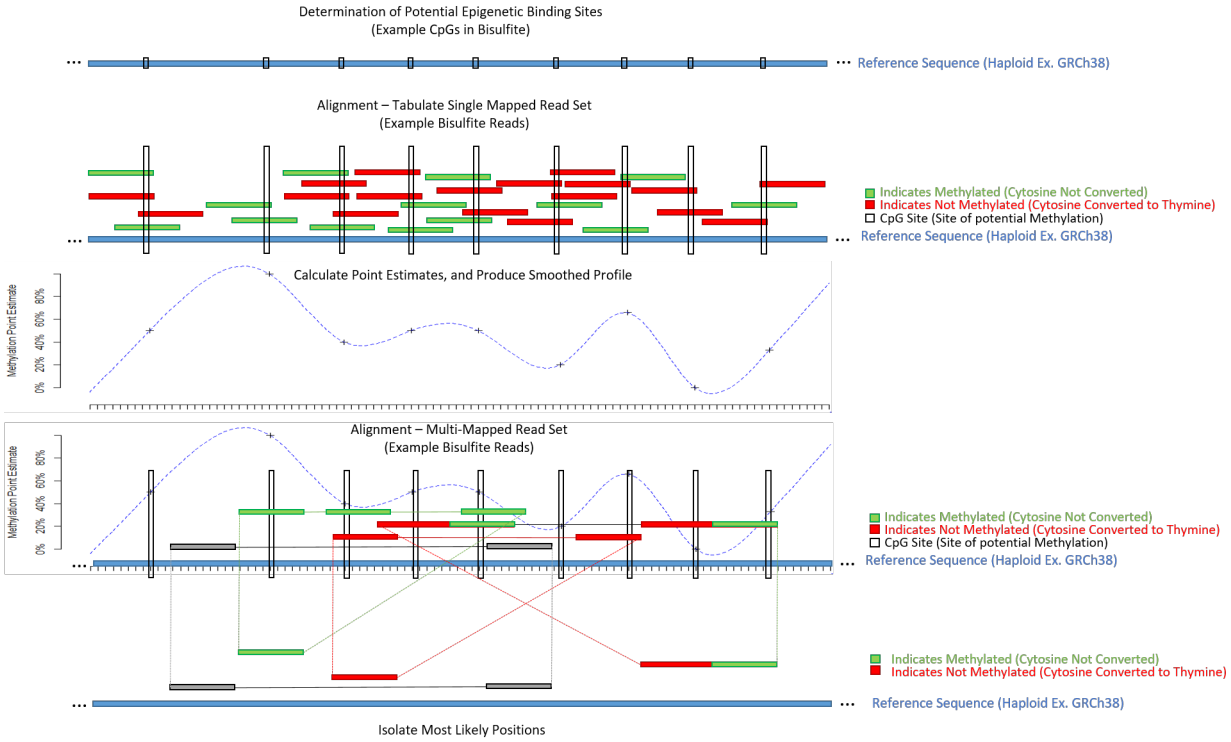


Figure 4.6: Illustration of the Multi-Mapped Read location isolation procedure

4.5. Simulation of Methylation Profile Errors

In genetic methylation experiments, ground truth information about the true methylation proportion at each location is never available. This means that a smoothed estimate of observed data at each position may or may not actually be closer to the actually true proportion. To assess the way in which smoothing point estimates of methylation effects the overall estimates, and specifically in the presence of highly variable coverage across positions, a simulation study was undertaken.

The simulation procedure, implemented in C [Kernighan and Ritchie, 2006], includes a binomial random number generation procedure via the inverse probability transform, as well as a uniform generator for doubles between 0 and 1 with resolution determined by the inbuilt max_rand. The approach allows for user specification of a few settings.

- Reference Length (N) - The size of the reference along which the methylation patterns are observed.
- Number of Locations (n_l) - The number of locations along the reference where methylation could occur.
- Coverage Mean (μ_c) - The average coverage on (number of reads mapped over) each of the methylation locations along the reference.
- Coverage Variance (σ_c^2)- The variance of the coverage across all potential methylation locations.

The simulation procedure is as follows:

1. Sample without replacement n_l values from 1 to N, these are the methylation locations, l_i for $i \in \{1, \ldots, n_l\}$.

2. Generate vector of hits associated with each location
   $h_i \sim \text{Bin}(n = \mu_c \cdot \frac{1}{1-\frac{\sigma_c^2}{\mu_c}}, p = 1 - \frac{\sigma_c^2}{\mu_c})$.

3. Generate a true methylation probability
   $t_i \sim \text{U}(0, 1)$.

4. For the i^{th} position ($i \in \{1, 2, \ldots, n_l\}$):

   (a) Generate the number of hits that indicate methylation
   $m_i \sim \text{Bin}(n = h_i, p = t_i)$.

   (b) Compute the point estimate of methylation at each location
   $\widehat{t_i} = \frac{m_i}{h_i}$.

5. For each l_j for $j \in \{1, 2, \ldots, n_l\}$ compute smoothed estimate of methylation

$$\overline{t_j}(r) = \frac{\sum_{i=1}^{n_l} \hat{t}_i \cdot \mathbb{1}(|l_i - l_j| \leq r)}{\sum_{i=1}^{n_l} \mathbb{1}(|l_i - l_j| \leq r)}, \qquad r \in \{200, 1000\}$$

6. Across the full reference determine the error from the point estimate, and both smoothed estimates for each position:

   (a) Determine the squared error at each location for each method

   - $\epsilon_{pi} = (\hat{t}_i - t_i)^2$ for the point estimates,
   - $\epsilon_{s1i} = (\overline{t_i}(200) - t_i)^2$ for the local average with bandwidth 200, and
   - $\epsilon_{s2i} = (\overline{t_i}(1000) - t_i)^2$ for the local average with bandwidth 1000.

The results of the simulation are included in Table 4.1. In this simulation, the number of hits on the methylation locations are generated by the binomial distribution. Table 4.1 shows that the average error of the point estimates is generally lower than the smoothed estimates.

This highlights one of the limitations of this initial simulation study. That is, the methylation proportions which are nearby one another are assumed to be unrelated in the study. There is no inherent or induced correlation structure on these methylation proportions and therefore it is possible that two very unlike proportions are found next to one another.

The primary takeaway of the simulation is then that the point estimates of methylation tend to be more accurate in this case, while also maintaining a lower error variance. When the estimates are smoothed with the 200 and 1000 nucleotide sized windows the resulting estimates typically result in errors that are eight to ten times that of the point estimates.

Table 4.1: Simulation results for smooth point estimates of methylation

μ_c	σ_c^2	$\overline{\epsilon_{pi}}$	$\mathbf{SE}(\epsilon_{pi})$	ϵ_{s1i}	$\mathbf{SE}(\epsilon_{s1i})$	ϵ_{s2i}	$\mathbf{SE}(\epsilon_{s2i})$
15	10	0.0114	0.000336	0.0799	0.00592	0.0827	0.00561
15	30	0.0129	0.000623	0.0806	0.00588	0.0834	0.00561
15	60	0.0161	0.00155	0.0794	0.00596	0.0821	0.00568
30	10	0.00552	0.000136	0.0794	0.00576	0.0825	0.00556
30	30	0.00574	9.2e-05	0.079	0.00579	0.0819	0.00557
30	60	0.00586	9.5e-05	0.079	0.00581	0.0818	0.00556
45	10	0.00382	0.000124	0.0795	0.00581	0.0825	0.00562
45	30	0.00379	0.00013	0.0794	0.00579	0.0824	0.00559
45	60	0.00386	3.8e-05	0.0783	0.00562	0.0814	0.00548
60	10	0.00282	2.8e-05	0.0791	0.00582	0.0823	0.00564
60	30	0.00279	2e-05	0.0788	0.00561	0.0824	0.00553
60	60	0.00302	0.000159	0.0798	0.00587	0.0827	0.00562

4.6. Conclusions

This chapter contains information on some statistical approaches to modeling methylation data, and to differentiating particular regions depending on those models. First the CpG density across human chromosome 22 is analyzed, and techniques for using this density as the kernel of a smoother for methylation proportions are discussed.

A short explanation of the ability of the methylation proportions to assist in the deduplication of multimapped reads into their most likely positions is then discussed. This approach is useful and can potentially allow researchers to refine their prior sequence alignment results. Lastly, a simulation study of smoothed estimates of methylation with moving averages of different lengths is provided. The results indicate that when there is no correlation among true

methylation proportions across a reference, the smoothed point estimates typically produce larger errors on average. However, the variance of the errors for these smoothed estimates are also robust to changes in the variance of the coverage across the reference.

Several key directions are highlighted in this chapter. An analysis of a real dataset, and the production of a computer code for implementation of the multi-mapped reduction algorithm will be undertaken. The resulting code will be provided to researchers who would like to refine their bisulfite sequencing experiments. A more robust simulation which allows for the programmatic toggling of correlation structure on the true methylation patterns of the reference will also be performed. This will allow for the determination of the effect of certain correlation structures on smoothed estimate errors.

Additionally, the use of differential methylation patterns on alleles may be of assistance in diploid assembly and alignment, as opposed to haploid, which is the current state of the art. Another avenue for investigation is the ability of breakpoints in the methylation pattern, or sudden changes in the correlation structure of the methylation, to demarcate gene boundaries on a contiguous segment. This may greatly assist in the analysis and annotation of newly discovered sequences. The correlation structure of methylation for individuals may yield important insights into the ability of these breakpoints to determine significant gene boundaries.

Chapter 5

Conclusion & Future Directions

This dissertation describes a group of statistical methodologies for comparing some modern kinds of biological data. It is split into three chapters, each describing a different type of biological data, and introducing statistical analysis approaches for inspecting them in case studies. In Chapter 2, genetic sequences containing ordered sets of the base nucleotides are inspected by their Fourier series. An R-package for this type of analysis is constructed, and applied to a set of SARS-CoV-2 virus genomes. The results are compared with more traditional phylogeny building techniques, and an extension for statistical hypothesis testing is developed and discussed. A few extensions were also offered. One is the development of a statistical test for comparing the variance-covariance matrix of series data for specific application on ensembles of sequences. In addition, the post application of filters to screen for the most important portions of the power spectra on which to classify sequences was examined. The chapter concluded with a thorough analysis of the comparison of sequence distance methods through a current and relevant study of viral genomes.

In Chapter 3 a semi-parametric model and a non-parametric glycan rank proportion model are introduced and applied to the classification of active vs. latent tuberculosis cases. The results included a comparison to a previously utilized method known as PLS-DA, which was also implemented in an R-Package for analysis. The data in this chapter were collected by the Lu lab at UT Southwestern, and the modeling results on using these data for all models were displayed. It was shown that the semi-parametric model and the rank proportion model both improved classification accuracy over the original PLS-DA model.

In Chapter 4, the methylation of DNA is considered, and some previous software describing the analysis of these methylation patterns is discussed. The modeling assumptions for the methylation profile on a haploid reference are presented, and a simulation study showing

the sensitivity of moving averages to variance in coverage shows that the approach is affected by regional variations in coverage.

The procedures outlined here are unique to this work, and their implementations derived in this dissertation. The analysis and evaluation of these methods with biological data indicates that they may be useful in classification models under certain contexts.

5.1. Future Directions

The usage of a multiple-valued transform may allow for more accurate characterization of the periodic trends exhibited by discretely indexed genomic signals (DNA/RNA). Specifically, it may be possible to apply some variant of the Walsh-Fourier transform which allows for variable length sequences, that is not constrained to signals of lengths which are powers of two. Careful statistical consideration should be applied when determining which procedure is the most valid, but the Levy-Chrestenson transform may be a viable candidate [Lévy, 1943, Chrestenson et al., 1955]. In the analysis of the genomic Fourier coefficients, another pressing direction is the application of the statistical hypothesis test discussed in Chapter 2 to the genomic Fourier coefficients. This analysis will include a discussion of the robustness of the statistical hypothesis test to violations of the assumption of second-order stationarity. An important next step is the merging of the Fourier Analysis techniques discussed in the first section with the methylation profiles discussed in the second session with regards to smoothing procedures. Another direction for analysis is a study of the sensitivity of the coefficients to non-identifiable locations, and other forms of missing data. A thorough investigation of different scaling procedures for scaling the Fourier Coefficients out to be equal length for all signals would also be a valuable pursuit in the future.

Secondly, In terms of the antibody glycosylation cellular protein study described in section three, extensions to the glycan rank probability model for using the empirical distributions of the actual ratios of proportions in the data will be investigated. This in addition to further modeling and the investigation of applications of these models to different kinds of

protein datasets will be undertaken.

Lastly, methylation smoothing as described in this dissertation is still being analyzed, and there are plans for including a more thorough analysis of *BSmooth* in addition to other smoothing techniques for estimating methylation profiles. It was previously shown that the local likelihood approach to smoothing the methylation profile might be only marginally better than a simple windowed average approach 4 [Wu et al., 2015]. The susceptibility of smoothing methods to large changes in sample size for proportion estimates across the length of the genome was also investigated. In the future, this will be assessed in relation to different types of bisulfite sequencing experiments. Further analysis will involve the derivation of an algorithm for effectively taking advantage of the methylation proportion data to isolate the proper locations of multi-mapped reads in software that aligns reads, using subsets of the traditional four nucleotides. Finally, some consideration to the estimation of dual methylation proportions in regards to the diploid nature of the human genome will be considered.

Appendix A

Genomic Fourier Coefficients: Tables & Figures

Table A.1: Kinds of post-MSA Phylogenetic Comparisons

Method Name	**Abbr.**	**Authors**	**Description**
Transversions Count	TV	-	A count of the number of transversions, that is A or G to C or T mismatches in aligned sequences
Galatier-Gouy	GG96	Nicolas Galtier Manolo Gouy	A substitution model based approach which allows a time-varying coefficient for specific nucleotides. [Galtier and Gouy, 1995]
Transitions Count	TS	-	A count of the number of transitions counted, that is A to G or C to T mismatches in aligned sequences
Barry-Hartigan	BH87	Daniel Barry John A Hartigan	A distance that is asymetric, and treats each possible transition/transversion differently [Barry and Hartigan, 1987]
Paralinear distance	paralin	James A. Lake	Uses an assymetric substitution model such as that used in the BH87 distance. [Lake, 1994]

MEGA	TN93	Koichiro Tamura and Masatoshi Nei	Specific kinds of transitions and transversions are weighted differently in this approach [Kumar et al., 1994] [Tamura and Nei, 1993]
Jukes-Cantor	JC69	Thomas C. Jukes Charles R. Cantor	An early technique which uses the mutations to determine the distance by a substitution model, this is a function of the P-distance (raw/N) [Jukes et al., 1969]
Generalized Jukes-Cantor	F81	Joseph Felsenstein	A theoretical technique that is the generalization of the Jukes-Cantor procedure in that it allows for different rates for each possible mismatch. [Felsenstein, 1981]
P-distance	raw	-	Proportion of sites that differ (sum of transversions and transitions) same as N.
P-distance	N	-	Proportion of sites that differ (sum of transversions and transitions) same as raw.
Kimura 3 parameteri distance	K81	Motoo Kimura	Jukes Cantor distance updated to allow different rates of 2 kinds transversions and transitions. [Kimura, 1981]
Kimura 2 parameter distance	K80	Motoo Kimura	Jukes Cantor distance updated to allow different rates of transversions and transitions. [Kimura, 1980]

Phylip	F84	Joseph Felsenstein and Gary A Churchill	A Markov-model based approach to phylogenetic distance calculations based on likely substitution states. [Felsenstein and Churchill, 1996]
Tamura (1992)	T92	Koichiro Tamura	assumes bases occur with unequal proportions, and allows variable rates in modeling procedure [Tamura, 1992].
K-mer distance	fivermers	Multiple	Euclidean distance computed between frequency vectors for the k-mers where $k \in \{1, 2, 3, 4, 5\}$.
Log Determinant	logdet	Peter J Lockhart and Michael A Steel	Another generalization of the BH procedure. [Lockhart et al., 1994]
DFT Power Spectral Distances	DFTPS	Changchuan Yin & Stephen S-T. Yau	Fourier transforms taken in the described manner produce power spectra which are scaled prior to the computation of Euclidean distances [Yin, 2020].
Insertion Deletion Count	indel	-	Counts mismatches among aligned sequences where one has a gap character (-).
Chained Insertion Deletion Count	indelblock	-	Counts mismatches among aligned sequences where one has one or more gap character(s) (*-*).

Note, most of the data in Table A.1 comes from the thorough documentation of the `ape` *package in R [Paradis and Schliep, 2019] [R Core Team, 2021].*

Appendix B

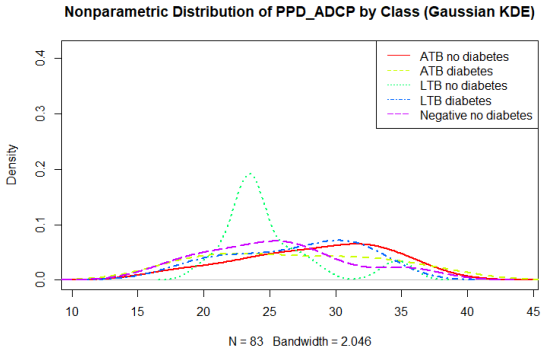
Modeling Compositional Data: Tables & Figures

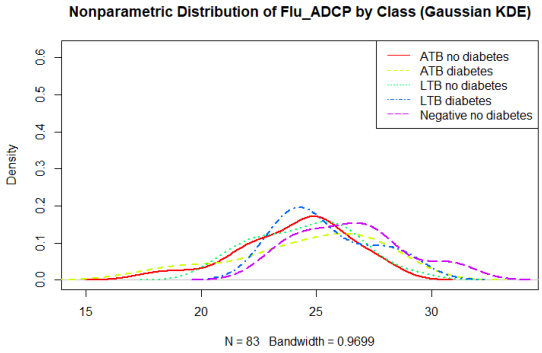
Table B.1: Non-Compositional Data (Non-Glycan Data)

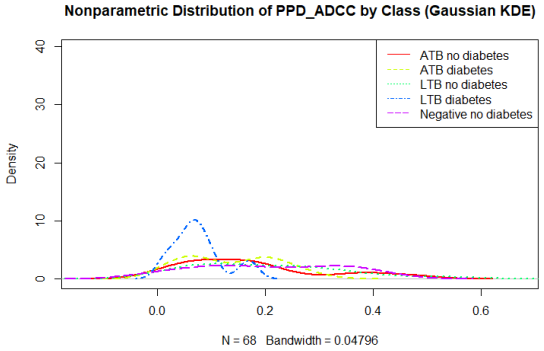
Variable Name	**Description**
PPD_ADCP	Mycobacterium tuberculosis purified protein derivative (PPD) specific antibody dependent cellular phagocytosis (ADCP) - Percent bead-positive frequency (MFI/10).
Flu_ADCP	Influenza Hemaglutinnin (HA)-specific ADCP.
PPD_ADCC	PPD-specific antibody dependent cellular cytotoxicity (percentage dead target cells per total target cells, pulsed less unpulsed cells).
IgG ELISA	An Enzyme-Linked Immunosorbent Assay (ELISA) for detecting and quantifying the amount of Immunoglobulin G in a tuberculosis patients peripheral blood sample.

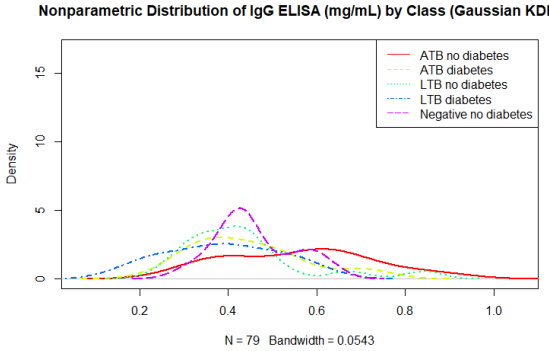
PPD Luminex AUC	A Luminex assay is another kind of procedure that can allow for the quantification of a particular protein, these particular measurements are made at three different concentrations of the IgG protein (0.1, 0.01, and 0.001 miligrams per mililiter) and reported as the total area under the curve (AUC) for the three (the sum of the three measurements).
Flu Luminex AUC	A Luminex Assay for Influenze Hemaglutinnin (HA) reported in the same style as the other Luminex AUCs.
RSV Luminex AUC	A Luminex assay for respiratory syncytial virus protein G, and F specific IgG titers reported in the same style as the other Luminex AUCs.
ESAT6 Luminex AUC	Mycobacterium tuberculosis ESAT-6 specific IgG titers reported in the same style as the other Luminex AUCs.
CFP10 Luminex AUC	Mycobacterium tuberculosis CFP10 specific IgG titers reported in the same style as the other Luminex AUCs.
Ag85AB Luminex AUC	Mycobacterium tuberculosis AG85AB specific IgG titers reported in the same style as the other Luminex AUCs.
GroES Luminex AUC	Mycobacterium tuberculosis Gro-ES specific IgG titers reported in the same style as the other Luminex AUCs.

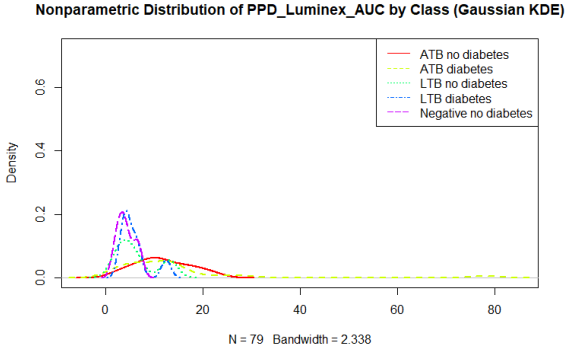
HspX Luminex AUC	Mycobacterium tuberculosis HspX specific IgG titers reported in the same style as the other Luminex AUCs.
Cell wall Luminex AUC	Mycobacterium tuberculosis Cell wall protein mixture specific IgG titers reported in the same style as the other Luminex AUCs.
PstS1 Luminex AUC	Mycobacterium tuberculosis PstS1-specific IgG titers reported in the same style as the other Luminex AUCs.
Cell Wall Lipids Luminex AUC	Mycobacterium tuberculosis cell wall lipid-specific IgG titers reported in the same style as the other Luminex AUCs.
Total Lipids Luminex AUC	Mycobacterium tuberculosis lipid-specific IgG titers reported in the same style as the other Luminex AUCs.
Lipomannan Luminex AUC	Mycobacterium tuberculosis lipomannan-specific IgG titers reported in the same style as the other Luminex AUCs.
ManLAM Luminex AUC	Mycobacterium tuberculosis mannose-capped lipomannan specific IgG titers reported in the same style as the other Luminex AUCs.
UT12 AUC	Mycobacterium tuberculosis growth of donor 12 macrophages over the course of 5 days.

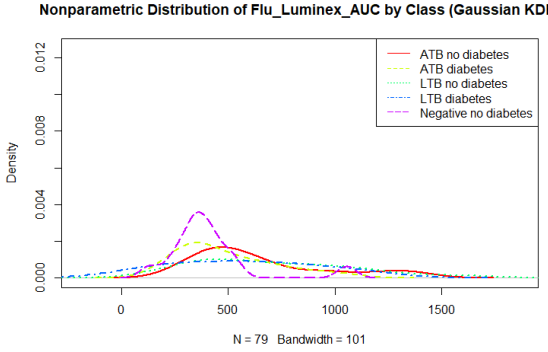

Nonparametric Distribution of PPD_ADCP by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 83 Bandwidth = 2.046

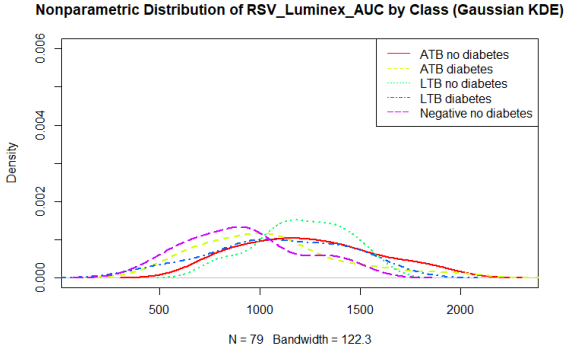

Nonparametric Distribution of Flu_ADCP by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 83 Bandwidth = 0.9699

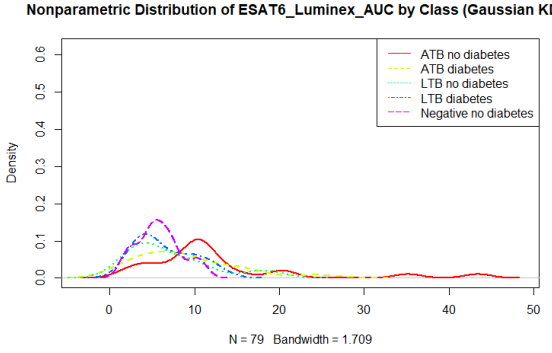

Nonparametric Distribution of PPD_ADCC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 68 Bandwidth = 0.04796

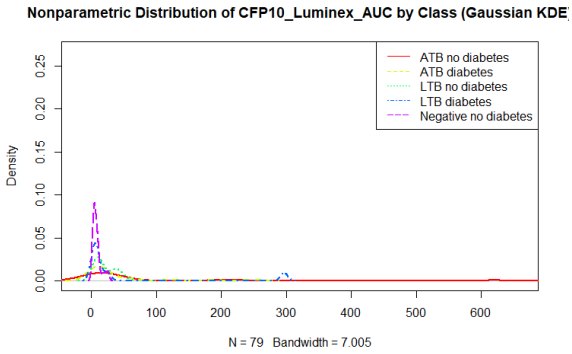

Nonparametric Distribution of IgG ELISA (mg/mL) by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 0.0543

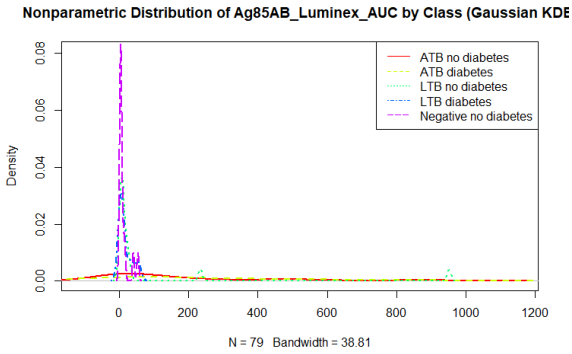

Nonparametric Distribution of PPD_Luminex_AUC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 2.338

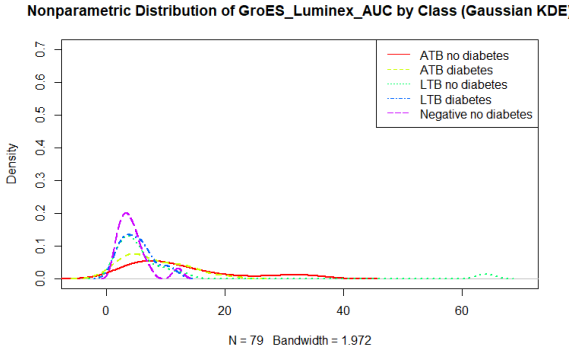

Nonparametric Distribution of Flu_Luminex_AUC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 101

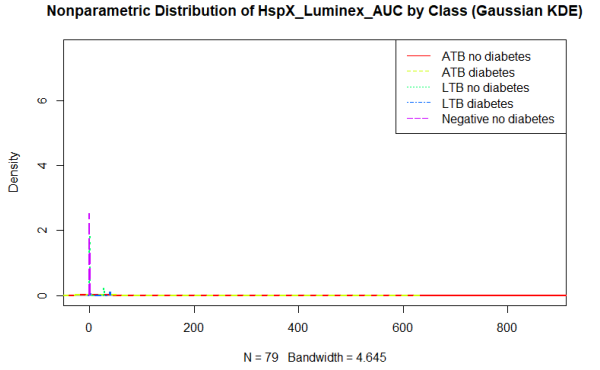

Nonparametric Distribution of RSV_Luminex_AUC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 122.3

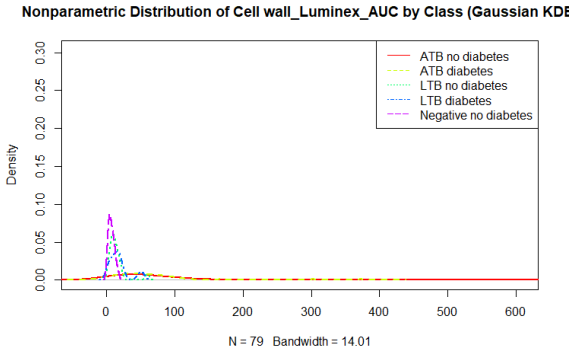

Nonparametric Distribution of ESAT6_Luminex_AUC by Class (Gaussian KDE
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 1.709

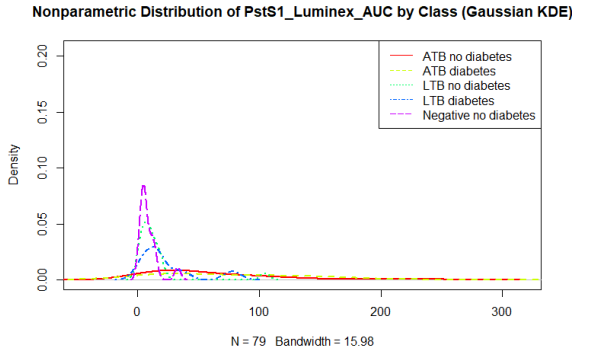

Nonparametric Distribution of CFP10_Luminex_AUC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 7.005


Nonparametric Distribution of Ag85AB_Luminex_AUC by Class (Gaussian KDE
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 38.81


Nonparametric Distribution of GroES_Luminex_AUC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 1.972


Nonparametric Distribution of HspX_Luminex_AUC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 4.645


Nonparametric Distribution of Cell wall_Luminex_AUC by Class (Gaussian KDE
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 14.01


Nonparametric Distribution of PstS1_Luminex_AUC by Class (Gaussian KDE)
ATB no diabetes
ATB diabetes
LTB no diabetes
LTB diabetes
Negative no diabetes
Density
N = 79 Bandwidth = 15.98

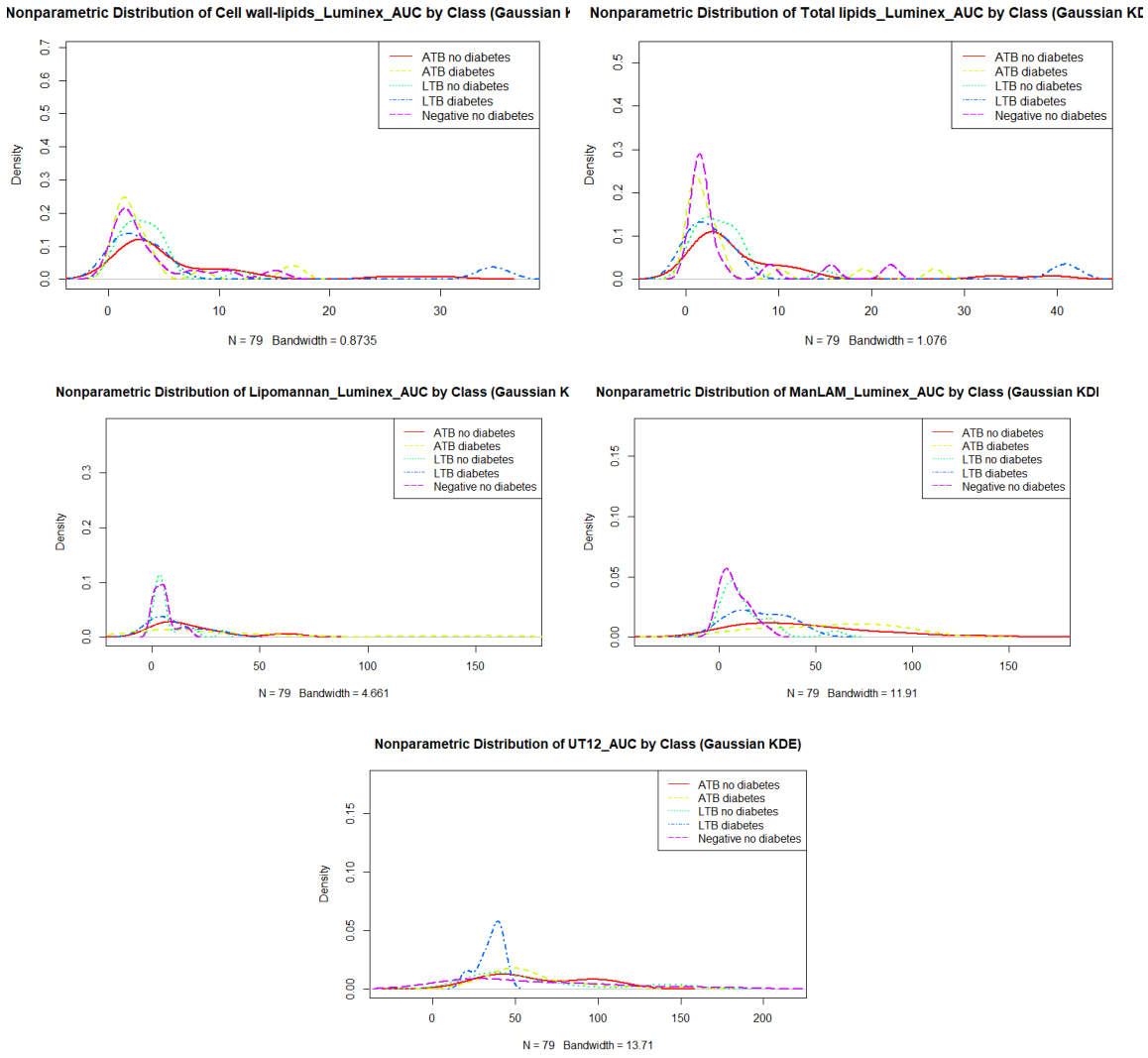


Figure B.1: Plots of non-parametric variables within five classes

Table B.2: Demographics for 2021 Tuberculosis Dataset

	Total	Kind of Tuberculosis			Significance Testing (Chi-Square)
	Count (Col Percent)	Active (Row Percent)	Latent (Row Percent)	Negative (Row Percent)	Chi-Square (p-value) Sig?
Total					
All Patients	83	43 (51.80723%)	25 (30.12048%)	15 (18.07229%)	
Sex					
Female	41 (49.398%)	20 (48.780%)	13 (31.707%)	8 (19.512%)	
Male	42 (50.602%)	23 (54.762%)	12 (28.571%)	7 (16.667%)	0.304 (0.8590033)
Diabetic (2010)					
Diabetic	25 (30.120%)	18 (72.000%)	7 (28.000%)	0 (0.000%)	
Not Diabetic	58 (69.880%)	25 (43.103%)	18 (31.034%)	15 (25.862%)	9.335 (0.009397337) **
TSPOT (stimulated w/ saline)					
0 Spots	12 (14.458%)	0 (0.000%)	12 (100.000%)	0 (0.000%)	
1 Spot	5 (6.024%)	0 (0.000%)	5 (100.000%)	0 (0.000%)	
2 Spots	6 (7.229%)	0 (0.000%)	6 (100.000%)	0 (0.000%)	
4 Spots	1 (1.205%)	0 (0.000%)	1 (100.000%)	0 (0.000%)	
5 Spots	1 (1.205%)	0 (0.000%)	1 (100.000%)	0 (0.000%)	
Unknown	58 (69.880%)	43 (74.138%)	0 (0.000%)	15 (25.862%)	83.000 (1.293034e-13) **
TSPOT (stimulated w/ CFP10					
10 Spots	1 (1.205%)	0 (0.000%)	1 (100.000%)	0 (0.000%)	
13 Spots	3 (3.614%)	0 (0.000%)	3 (100.000%)	0 (0.000%)	
14 Spots	2 (2.410%)	0 (0.000%)	2 (100.000%)	0 (0.000%)	
20 Spots	17 (20.482%)	0 (0.000%)	17 (100.000%)	0 (0.000%)	
9 Spots	2 (2.410%)	0 (0.000%)	2 (100.000%)	0 (0.000%)	
Unknown	58 (69.880%)	43 (74.138%)	0 (0.000%)	15 (25.862%)	83.000 (1.293034e-13) **
TSPOT (stimulated w/ ESAT6					
0 Spots	2 (2.410%)	0 (0.000%)	2 (100.000%)	0 (0.000%)	
11 Spots	2 (2.410%)	0 (0.000%)	2 (100.000%)	0 (0.000%)	
13 Spots	2 (2.410%)	0 (0.000%)	2 (100.000%)	0 (0.000%)	
15 Spots	3 (3.614%)	0 (0.000%)	3 (100.000%)	0 (0.000%)	
16 Spots	1 (1.205%)	0 (0.000%)	1 (100.000%)	0 (0.000%)	
2 Spots	2 (2.410%)	0 (0.000%)	2 (100.000%)	0 (0.000%)	
20 Spots	7 (8.434%)	0 (0.000%)	7 (100.000%)	0 (0.000%)	
3 Spots	3 (3.614%)	0 (0.000%)	3 (100.000%)	0 (0.000%)	
4 Spots	2 (2.410%)	0 (0.000%)	2 (100.000%)	0 (0.000%)	
5 Spots	1 (1.205%)	0 (0.000%)	1 (100.000%)	0 (0.000%)	
Unknown	58 (69.880%)	43 (74.138%)	0 (0.000%)	15 (25.862%)	83.000 (1.208326e-09) **
TSPOT Interpretation					
Unknown	43 (51.807%)	43 (100.000%)	0 (0.000%)	0 (0.000%)	
Negative	15 (18.072%)	0 (0.000%)	0 (0.000%)	15 (100.000%)	
Positive	25 (30.120%)	0 (0.000%)	25 (100.000%)	0 (0.000%)	166.000 (7.548094e-35) **
Recieved Bacille Calmette-Guerin vaccination					
Unknown	18 (21.687%)	1 (5.556%)	2 (11.111%)	15 (83.333%)	
No	13 (15.663%)	10 (76.923%)	3 (23.077%)	0 (0.000%)	
Yes	52 (62.651%)	32 (61.538%)	20 (38.462%)	0 (0.000%)	67.707 (6.915114e-14) **
Place of Birth					
Mexico	36 (43.373%)	36 (100.000%)	0 (0.000%)	0 (0.000%)	
Unknown	40 (48.193%)	0 (0.000%)	25 (62.500%)	15 (37.500%)	
South Africa	1 (1.205%)	1 (100.000%)	0 (0.000%)	0 (0.000%)	
USA	6 (7.229%)	6 (100.000%)	0 (0.000%)	0 (0.000%)	83.000 (8.565786e-16) **
Smoking Status					
Unknown	17 (20.482%)	2 (11.765%)	0 (0.000%)	15 (88.235%)	
No	52 (62.651%)	30 (57.692%)	22 (42.308%)	0 (0.000%)	
Yes	14 (16.867%)	11 (78.571%)	3 (21.429%)	0 (0.000%)	73.816 (3.546616e-15) **
X-Ray Indicates Cavitary Disease					
Unknown	44 (53.012%)	4 (9.091%)	25 (56.818%)	15 (34.091%)	
No	25 (30.120%)	25 (100.000%)	0 (0.000%)	0 (0.000%)	
Yes	14 (16.867%)	14 (100.000%)	0 (0.000%)	0 (0.000%)	68.436 (4.854872e-14) **
Sputum Smear Scores					
1+ or lt 1 per field	6 (7.229%)	6 (100.000%)	0 (0.000%)	0 (0.000%)	
2+ or 1 to 10 per field	11 (13.253%)	11 (100.000%)	0 (0.000%)	0 (0.000%)	
3+ or gt 10 per field	18 (21.687%)	18 (100.000%)	0 (0.000%)	0 (0.000%)	
Unknown	40 (48.193%)	0 (0.000%)	25 (62.500%)	15 (37.500%)	
Negative	8 (9.639%)	8 (100.000%)	0 (0.000%)	0 (0.000%)	83.000 (1.214861e-14) **

Figure B.2: Partial Least Squares Tuberculosis/Diabetes Classification LV plot

Figure B.3: Partial Least Squares Tuberculosis Classification LV plot

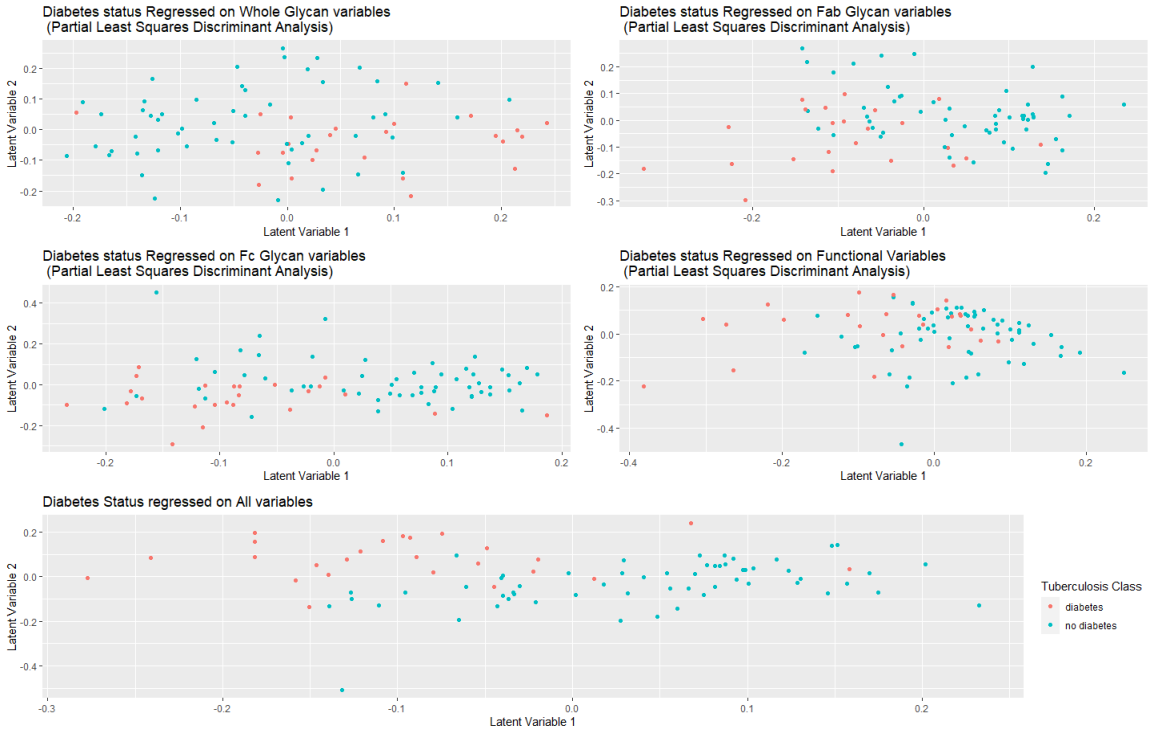


Figure B.4: Partial Least Squares Diabetes Classification LV plot

Figure B.5: Principal Components Plots Tuberculosis/Diabetes Biplot

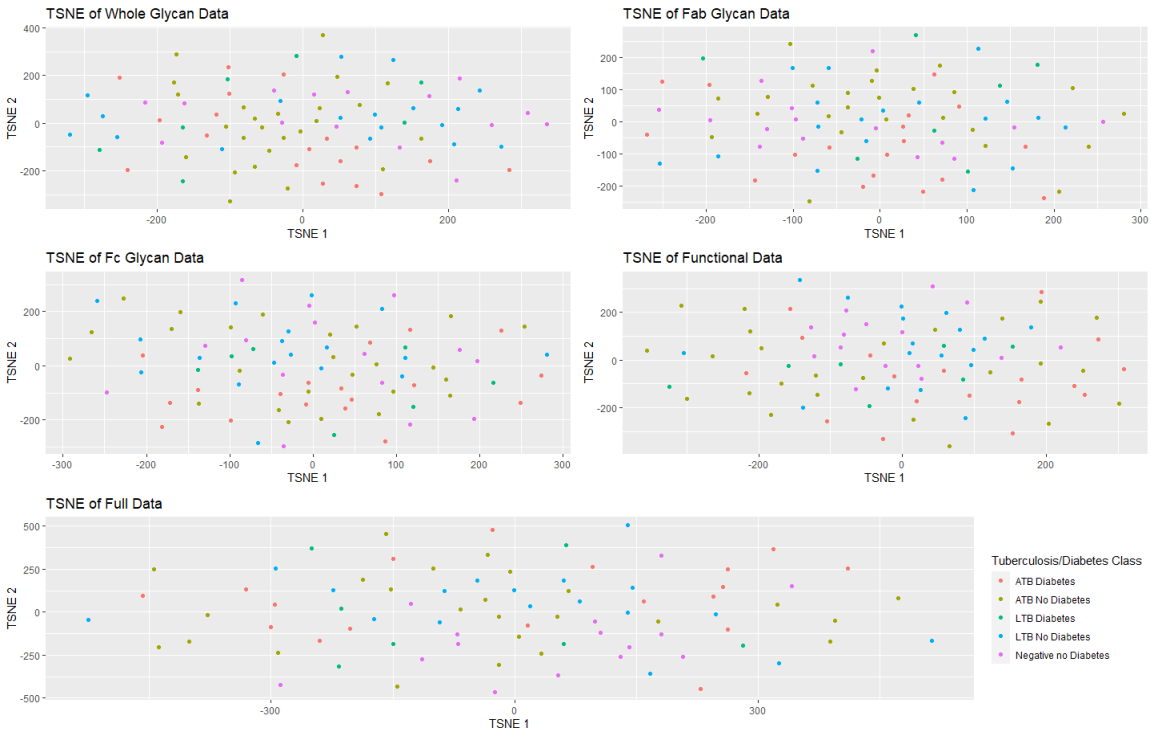


Figure B.6: TSNE Plots for Tuberculosis/Diabetes classification)

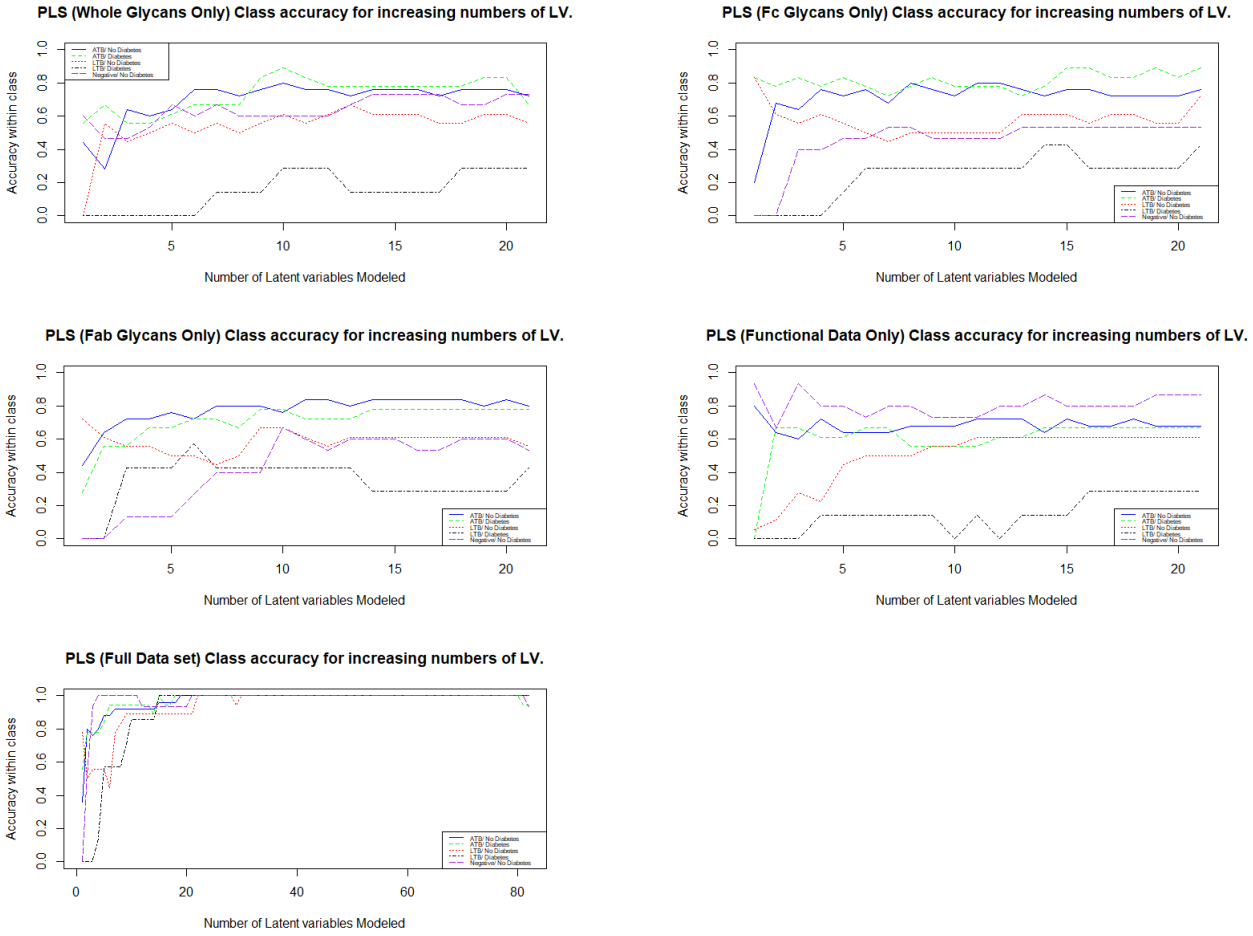


Figure B.7: Class Accuracy for PLS-DA classifier model

Table B.3: Semi-Parametric Confusion Matrix Tuberculosis Status (Full Data)

	True Class	Predicted Class		
		ATB	LTB	Neg
Nonparametric Score Only	ATB	33	9	1
	LTB	1	20	4
	Neg	0	1	14
Parametric (Multinomial) Score Only	ATB	31	7	5
	LTB	7	16	2
	Neg	4	4	7
Combined Semiparametric Score	ATB	35	2	1
	LTB	7	20	1
	Neg	0	2	12

Table B.4: Semi-Parametric Confusion Matrix Tuberculosis Status (Leave-one-out CV)

	True Class	Predicted Class		
		ATB	LTB	Neg
Nonparametric Score Only	ATB	32	10	1
	LTB	8	11	6
	Neg	0	10	5
Parametric (Multinomial) Score Only	ATB	31	7	5
	LTB	10	10	5
	Neg	5	4	6
Combined Semiparametric Score	ATB	34	8	0
	LTB	5	16	3
	Neg	3	9	2

Table B.5: Semi-Parametric Confusion Matrix Diabetes Status (full data)

	True Class	Predicted Class	
		DM	ND
Nonparametric Score Only	DM	15	10
	ND	3	55
Parametric (Multinomial) Score Only	DM	23	2
	ND	21	37
Combined Semiparametric Score	DM	21	13
	ND	4	42

Table B.6: Semi-Parametric Confusion Matrix Diabetes Status (Leave-One-Out CV)

	True Class	Predicted Class	
		DM	ND
Nonparametric Score Only	DM	8	17
	ND	15	43
Parametric (Multinomial) Score Only	DM	22	3
	ND	24	34
Combined Semiparametric Score	DM	20	5
	ND	21	34

REFERENCES

[Aitchison, 1982] Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2):139–160.

[Aitchison, 1983] Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika, 70(1):57–65.

[Aitchison, 1994] Aitchison, J. (1994). Principles of compositional data analysis. Lecture Notes-Monograph Series, pages 73–81.

[Aitchison and Greenacre, 2002] Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 51(4):375–392.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell. Garland Science, a member of the Taylor & Francis group, 29 West 35th Street, New York, NY 10001-2299.

[Allaire and Chollet, 2021] Allaire, J. and Chollet, F. (2021). keras: R Interface to 'Keras'. R package version 2.4.0.

[Allman et al., 2017] Allman, E. S., Rhodes, J. A., and Sullivant, S. (2017). Statistically consistent k-mer methods for phylogenetic tree reconstruction. Journal of Computational Biology, 24(2):153–171.

[Anastassiou, 2001] Anastassiou, D. (2001). Genomic signal processing. IEEE Signal Processing Magazine, 18(4):8–20.

[Armăşelu, 2017] Armăşelu, A. (2017). New spectral applications of the Fourier transforms in medicine, biological and biomedical fields. In Fourier Transforms - High-tech Application and Current Trends, pages 235–252. IntechOpen.

[Baerwald et al., 2016] Baerwald, M. R., Meek, M. H., Stephens, M. R., Nagarajan, R. P., Goodbla, A. M., Tomalty, K. M., Thorgaard, G. H., May, B., and Nichols, K. M. (2016). Migration-related phenotypic divergence is associated with epigenetic modifications in rainbow trout. Molecular Ecology, 25(8):1785–1800.

[Barry and Hartigan, 1987] Barry, D. and Hartigan, J. A. (1987). Statistical analysis of hominoid molecular evolution. Statistical Science, 2(2):191–207.

[Bell et al., 2012] Bell, J. T., Tsai, P.-C., Yang, T.-P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., et al. (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS genetics, 8(4):e1002629.

[Benoit et al., 2016] Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., and Lemaitre, C. (2016). Multiple comparative metagenomics using multiset k-mer counting. PeerJ Computer Science, 2:e94.

[Berosik et al., 2010] Berosik, S., Goudberg, J., Chhibber, A., Davidson, C., Felton, A., Fish, R., Gomi, T., Haraura, I., Hung, S., and et al., R. I. (2010). Innovative software, hardware, and consumable development for the new 3500 genetic analyzer system. Journal of Biomolecular Techniques: JBT, 21(3 Suppl):S27.

[Bird, 2002] Bird, A. (2002). Dna methylation patterns and epigenetic memory. Genes & development, 16(1):6–21.

[Blom et al., 1999] Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of molecular biology, 294(5):1351–1362.

[Bonasio et al., 2010] Bonasio, R., Tu, S., and Reinberg, D. (2010). Molecular signals of epigenetic states. science, 330(6004):612–616.

[Braden and Poljak, 1995] Braden, B. C. and Poljak, R. J. (1995). Structural features of the reactions between antibodies and protein antigens. FASEB J, 9(1):9–16.

[Breiman, 2001] Breiman, L. (2001). Random forests. Machine Learning, 45:5–32.

[Brereton and Lloyd, 2014] Brereton, R. G. and Lloyd, G. R. (2014). Partial least squares discriminant analysis: taking the magic away. Journal of Chemometrics, 28(4):213–225.

[Brieman et al., 1984] Brieman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). Classification and Regression Trees. Wadsworth, first edition edition.

[Brillinger, 2001] Brillinger, D. R. (2001). Time Series Data Analysis and Theory. Society for Industrial and Applied Mathematics.

[Centers for Disease Control and Prevention, 2021] Centers for Disease Control and Prevention (2021). Science brief: Emerging sars-cov-2 variants. Website.

[Chen et al., 2017] Chen, Y., Pal, B., Visvader, J. E., and Smyth, G. K. (2017). Differential methylation analysis of reduced representation bisulfite sequencing experiments using edger. F1000Research, 6.

[Chrestenson et al., 1955] Chrestenson, H. et al. (1955). A class of generalized walsh functions. Pacific Journal of Mathematics, 5(1):17–31.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support vector networks. Machine Learning, 20(3):273–297.

[Corum and Zimmer, 2020] Corum, J. and Zimmer, C. (2020). Bad news wrapped in protein: Inside the coronavirus genome. The New York Times.

[Dall'Aglio et al., 2018] Dall'Aglio, L., Muka, T., Cecil, C. A., Bramer, W. M., Verbiest, M. M., Nano, J., Hidalgo, A. C., Franco, O. H., and Tiemeier, H. (2018). The role of epigenetic modifications in neurodevelopmental disorders: A systematic review. Neuroscience & Biobehavioral Reviews, 94:17–30.

[Darst et al., 2010] Darst, R. P., Pardo, C. E., Ai, L., Brown, K. D., and Kladde, M. P. (2010). Bisulfite sequencing of dna. Current protocols in molecular biology, 91(1):7–9.

[De Leeuw and Mair, 2009] De Leeuw, J. and Mair, P. (2009). Multidimensional scaling using majorization: Smacof in r. Journal of statistical software, 31(1):1–30.

[Donaldson, 2016] Donaldson, J. (2016). tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE). R package version 0.1-3.

[Eddy et al., 2019] Eddy, A. C., Chapman, H., and George, E. M. (2019). Acute hypoxia and chronic ischemia induce differential total changes in placental epigenetic modifications. Reproductive Sciences, 26(6):766–773.

[Edwards et al., 2017] Edwards, J. R., Yarychkivska, O., Boulard, M., and Bestor, T. H. (2017). Dna methylation and dna methyltransferases. Epigenetics & chromatin, 10(1):1–10.

[Egozcue et al., 2003] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. Mathematical geology, 35(3):279–300.

[Elbe and Buckland-Merrett, 2017] Elbe, S. and Buckland-Merrett, G. (2017). Data, disease and diplomacy: Gisaid's innovative contribution to global health. Global Challenges, pages 33–46.

[Ewing et al., 1989] Ewing, A. G., Wallingford, R. A., and Olefirowicz, T. M. (1989). Capillary electrophoresis. Analytical Chemistry, 61(4):292A–303A.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. Journal of molecular evolution, 17(6):368–376.

[Felsenstein and Churchill, 1996] Felsenstein, J. and Churchill, G. A. (1996). A hidden markov model approach to variation among sites in rate of evolution. Molecular biology and evolution, 13(1):93–104.

[Filzmoser et al., 2018a] Filzmoser, P., Horn, K., and Templ, M. (2018a). Applied Compositional Data Analysis. Springer.

[Filzmoser et al., 2018b] Filzmoser, P., Hron, K., and Templ, M. (2018b). Applied compositional data analysis. Switzerland: Springer Nature.

[Flores et al., 2013] Flores, K. B., Wolschin, F., and Amdam, G. V. (2013). The role of methylation of dna in environmental adaptation.

[Frommer et al., 1992] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. Proceedings of the National Academy of Sciences, 89(5):1827–1831.

[Fuentes et al., 2006] Fuentes, A. R., Ginori, J. V. L., and Ãbalo, R. G. (2006). Detection of coding regions in large DNA sequences using the short time Fourier transform with reduced computational load.

[Galtier and Gouy, 1995] Galtier, N. and Gouy, M. (1995). Inferring phylogenies from dna sequences of unequal base compositions. Proceedings of the National Academy of Sciences, 92(24):11317–11321.

[Gebrehiwot et al., 2018] Gebrehiwot, A. G., Melka, D. S., Kassaye, Y. M., Rehan, I. F., Rangappa, S., Hinou, H., Kamiyama, T., and Nishimura, S.-I. (2018). Healthy human serum n-glycan profiling reveals the influence of ethnic variation on the identified cancer-relevant glycan biomarkers. PLoS One, 13(12):e0209515.

[Geladi, 1988] Geladi, P. (1988). Notes on the history and nature of partial least squares (pls) modelling. Journal of Chemometrics, 2(4):231–246.

[Geladi and Kowalski, 1986] Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. Analytica chimica acta, 185:1–17.

[Genereux et al., 2005] Genereux, D. P., Miner, B. E., Bergstrom, C. T., and Laird, C. D. (2005). A population-epigenetic model to infer site-specific methylation rates from double-stranded dna methylation patterns. Proceedings of the National Academy of Sciences, 102(16):5802–5807.

[Global Tuberculosis Report 2020, 2020] Global Tuberculosis Report 2020 (2020). Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO.

[Goodman, 1965] Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. Technometrics, 7(2):247–254.

[Gordon et al., 1988] Gordon, M. J., Huang, X., Pentoney, S. L., and Zare, R. N. (1988). Capillary electrophoresis. Science, 242(4876):224–228.

[Grenfell et al., 2001] Grenfell, B. T., Bjorndstad, O. N., and Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. Nature, 414(6865):716.

[Gromski et al., 2015] Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., and Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. Analytica chimica acta, 879:10–23.

[Grossman and Colburn, 2012] Grossman, P. D. and Colburn, J. C. (2012). Capillary electrophoresis: Theory and practice. Academic Press.

[Gudelj et al., 2018] Gudelj, I., Lauc, G., and Pezer, M. (2018). Immunoglobulin g glycosylation in aging and diseases. Cellular immunology, 333:65–79.

[Gunn and Alter, 2016] Gunn, B. M. and Alter, G. (2016). Modulating antibody functionality in infectious disease and vaccination. Trends in molecular medicine, 22(11):969–982.

[Hansen et al., 2012] Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biology, 13(R83).

[Hastie et al., 2008] Hastie, T., Tibshirani, R., and Friedman, J. (2008). The elements of statistical learning: data mining, inference and prediction. Springer New York.

[Heideman et al., 1984] Heideman, M., Johnson, D., and Burrus, C. (1984). Gauss and the history of the fast Fourier transform. IEEE ASSP Magazine, 1(4):14–21.

[Henikoff and Shilatifard, 2011] Henikoff, S. and Shilatifard, A. (2011). Histone modification: cause or cog? Trends in Genetics, 27(10):389–396.

[Hinton and Roweis, 2002] Hinton, G. E. and Roweis, S. (2002). Stochastic neighbor embedding. Advances in Neural Information Processing Systems, 15:857–864.

[Hirabayashi, 2008] Hirabayashi, J. (2008). Glycan Profiling, pages 56–59. Springer Japan, Tokyo.

[Hopfield, 1982] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Procedings of the National Academy of Sciences USA, 79(8):2554–2558.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24:417–441.

[Hotelling, 1935] Hotelling, H. (1935). The most predictable criterion. Journal of educational Psychology, 26(2):139.

[Hron et al., 2010] Hron, K., Templ, M., and Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis, 54(12):3095–3107.

[Jaynes, 2003] Jaynes, E. T. (2003). Probability theory: The logic of science. Cambridge university press.

[Johnson and Goody, 2011] Johnson, K. A. and Goody, R. S. (2011). The original Michaelis constant: Translation of the 1913 Michaelis–menten paper. Biochemistry, 50(39):8264–8269.

[Jones and Liang, 2009] Jones, P. A. and Liang, G. (2009). Rethinking how dna methylation patterns are maintained. Nature Reviews Genetics, 10(11):805–811.

[Jukes et al., 1969] Jukes, T. H., Cantor, C. R., et al. (1969). Evolution of protein molecules. Mammalian protein metabolism, 3:21–132.

[Karlić et al., 2010] Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. Proceedings of the National Academy of Sciences, 107(7):2926–2931.

[Kassambara and Mundt, 2020] Kassambara, A. and Mundt, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7.

[Katoh et al., 2002] Katoh, K., Misawa, K., ichi Kuma, K., and Miyata, T. (2002). MaFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 30(14):3059–3066. PMC135756[pmcid].

[Kawata, 1966] Kawata, T. (1966). On the Fourier series of a stationary stochastic process. Z. Wahrscheinlichkeitstheorie verw Gebiete, 6:224–245.

[Kernighan and Ritchie, 2006] Kernighan, B. W. and Ritchie, D. M. (2006). The C programming language.

[Kettenring, 1971] Kettenring, J. R. (1971). Canonical analysis of several sets of variables. Biometrika, 58(3):433–451.

[Kim et al., 2015] Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. Nature methods, 12(4):357–360.

[Kim et al., 2019] Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nature Biotechnology, 37(8):907–915.

[Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution, 16(2):111–120.

[Kimura, 1981] Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences, 78(1):454–458.

[Konopka, 2020] Konopka, T. (2020). umap: Uniform Manifold Approximation and Projection. R package version 0.2.7.0.

[Krueger et al., 2012] Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). Dna methylome analysis using short bisulfite sequencing data. Nature methods, 9(2):145–151.

[Kumar et al., 1994] Kumar, S., Tamura, K., and Nei, M. (1994). Mega: molecular evolutionary genetics analysis software for microcomputers. Bioinformatics, 10(2):189–191.

[Lake, 1994] Lake, J. A. (1994). Reconstructing evolutionary trees from dna and protein sequences: paralinear distances. Proceedings of the National Academy of Sciences, 91(4):1455–1459.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. Nature methods, 9(4):357.

[Larkin et al., 2007] Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., and et al., R. L. (2007). Clustal w and clustal x version 2.0. bioinformatics, 23(21):2947–2948.

[Lattová et al., 2020] Lattová, E., Skřičková, J., Hausnerová, J., Frola, L., Křen, L., Ihnatová, I., Zdráhal, Z., Bryant, J., and Popovič, M. (2020). N-glycan profiling of lung adenocarcinoma in patients at different stages of disease. Modern Pathology, 33(6):1146–1156.

[Leenen et al., 2016] Leenen, F. A., Muller, C. P., and Turner, J. D. (2016). Dna methylation: conducting the orchestra from exposure to phenotype? Clinical epigenetics, 8(1):1–15.

[Lévy, 1943] Lévy, P. (1943). Sur une généralisation des fonctions orthogonales de m. rademacher. Commentarii Mathematici Helvetici, 16(1):146–152.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14):1754–1760.

[Liang et al., 2021] Liang, Z.-Z., Zhu, R.-M., Li, Y.-L., Jiang, H.-M., Li, R.-B., Wang, Q., Tang, L.-Y., and Ren, Z.-F. (2021). Differential epigenetic profiles induced by sodium selenite in breast cancer cells. Journal of Trace Elements in Medicine and Biology, 64:126677.

[Lipman and Pearson, 1985] Lipman, D. and Pearson, W. (1985). Rapid and sensitive protein similarity searches. Science, 227(4693):1435–1441.

[Liu et al., 1999] Liu, L., Scolnick, D. M., Trievel, R. C., Zhang, H. B., Marmorstein, R., Halazonetis, T. D., and Berger, S. L. (1999). p53 sites acetylated in vitro by pcaf and p300 are acetylated in vivo in response to dna damage. Molecular and cellular biology, 19(2):1202–1209.

[Loader, 1999] Loader, C. (1999). Local Regression and Likelihood. Springer-Verlag New York Inc., 175 Fifth Avenue, New York, NY 10010, USA.

[Lockhart et al., 1994] Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Molecular biology and evolution, 11(4):605–612.

[Lu et al., 2018] Lu, G., Crihfield, C. L., Gattu, S., Veltri, L. M., and Holland, L. A. (2018). Capillary electrophoresis separations of glycans. Chemical reviews, 118(17):7867–7885.

[Lu et al., 2021] Lu, L., Miles, J., and Lu, P. (2021). Private Communication.

[Lu et al., 2016] Lu, L. L., Chung, A. W., Rosebrock, T. R., Ghebremichael, M., Yu, W. H., Grace, P. S., Schoen, M. K., Tafesse, F., Martin, C., and et al., V. L. (2016). A functional role for antibodies in tuberculosis. Cell, 167(2):433–443.

[Lu et al., 2020] Lu, L. L., Das, J., Grace, P. S., Fortune, S. M., Restrepo, B. I., and Alter, G. (2020). Antibody fc glycosylation discriminates between latent and active tuberculosis. The Journal of Infectious Diseases, 222(12):2093–2102.

[Lux and Nimmerjahn, 2011] Lux, A. and Nimmerjahn, F. (2011). Impact of differential glycosylation on igg activity. Crossroads between Innate and Adaptive Immunity III, pages 113–124.

[Marcais and Kingsford, 2012] Marcais, G. and Kingsford, C. (2012). Jellyfish: A fast k-mer counter. Tutorialis e Manuais, 1:1–8.

[Mardia, 1978] Mardia, K. V. (1978). Some properties of clasical multi-dimesional scaling. Communications in Statistics-Theory and Methods, 7(13):1233–1241.

[Marsit et al., 2006] Marsit, C. J., Houseman, E. A., Christensen, B. C., Eddy, K., Bueno, R., Sugarbaker, D. J., Nelson, H. H., Karagas, M. R., and Kelsey, K. T. (2006). Examination of a CpG island methylator phenotype and implications of methylation profiles in solid tumors. Cancer research, 66(21):10621–10629.

[MATLAB, 2020] MATLAB (2020). 9.8.0.1417392 (R2020a) Update 4 (R2020a). The MathWorks Inc., Natick, Massachusetts.

[McCartney et al., 2018] McCartney, D. L., Stevenson, A. J., Hillary, R. F., Walker, R. M., Bermingham, M. L., Morris, S. W., Clarke, T.-K., Campbell, A., Murray, A. D., Whalley, H. C., et al. (2018). Epigenetic signatures of starting and stopping smoking. EBioMedicine, 37:214–220.

[Mechref et al., 2012] Mechref, Y., Hu, Y., Garcia, A., Zhou, S., Desantos-Garcia, J. L., and Hussein, A. (2012). Defining putative glycan cancer biomarkers by ms. Bioanalysis, 4(20):2457–2469.

[Medvedeva et al., 2015] Medvedeva, Y. A., Lennartsson, A., Ehsani, R., Kulakovskiy, I. V., Vorontsov, I. E., Panahandeh, P., Khimulya, G., Kasukawa, T., Drabløs, F., Consortium, F., et al. (2015). Epifactors: a comprehensive database of human epigenetic factors and complexes. Database, 2015.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics, 11(1):31–46.

[Michaelis and Menten, 1913] Michaelis, L. and Menten, M. L. (1913). Die kinetik der invertinwirkung. Biochemische Zeitschrift, 49:333 – 369.

[Mittermayr et al., 2013] Mittermayr, S., Bones, J., and Guttman, A. (2013). Unraveling the glyco-puzzle: glycan structure identification by capillary electrophoresis. Analytical chemistry, 85(9):4228–4238.

[Morgan, 2021] Morgan, M. (2021). BiocManager: Access the Bioconductor Project Package Repository. R package version 1.30.16.

[Nagel et al., 2001] Nagel, H. R., Granum, E., and Musaeus, P. (2001). Methods for visual mining of data in virtual reality. In Proceedings of the International Workshop on Visual Data Mining, pages 13–27.

[Nakagawa et al., 2007] Nakagawa, H., Hato, M., Takegawa, Y., Deguchi, K., Ito, H., Takahata, M., Iwasaki, N., Minami, A., and Nishimura, S.-I. (2007). Detection of altered n-glycan profiles in whole serum from rheumatoid arthritis patients. Journal of Chromatography B, 853(1-2):133–137.

[North, 1963] North, D. O. (1963). An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems. Proceedings of the IEEE, 51(7):1016–1027.

[Nunes et al., 2012] Nunes, C. A., Freitas, M. P., Pinheiro, A. C. M., and Bastos, S. C. (2012). Chemoface: a novel free user-friendly interface for chemometrics. Journal of the Brazilian Chemical Society, 23:2003–2010.

[O'Brien et al., 2006] O'Brien, M. J., Yang, S., Mack, C., Xu, H., Huang, C. S., Mulcahy, E., Amorosino, M., and Farraye, F. A. (2006). Comparison of microsatellite instability, cpg island methylation phenotype, braf and kras status in serrated polyps and traditional adenomas indicates separate pathways to distinct colorectal carcinoma end points. The American journal of surgical pathology, 30(12):1491–1501.

[Oppenheim et al., 1997] Oppenheim, A. V., Willisky, A. S., and Nawab, S. H. (1997). Signals & Systems 2nd Edition. Pearson Education, Upper Saddle River, New Jersey 07458.

[Paradis and Schliep, 2019] Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35:526–528.

[Park and Wu, 2016] Park, Y. and Wu, H. (2016). Differential methylation analysis for bs-seq data under general experimental design. Bioinformatics, 32(10):1446–1453.

[Pawlowsky-Glahn and Buccianti, 2011] Pawlowsky-Glahn, V. and Buccianti, A. (2011). Compositional data analysis: Theory and applications. John Wiley & Sons.

[Pawlowsky-Glahn and Egozcue, 2006] Pawlowsky-Glahn, V. and Egozcue, J. J. (2006). Compositional data and their analysis: an introduction. Geological Society, London, Special Publications, 264(1):1–10.

[Pawlowsky-Glahn et al., 2015] Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. John Wiley & Sons.

[Pérez-Enciso and Tenenhaus, 2003] Pérez-Enciso, M. and Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach. Human genetics, 112(5-6):581–592.

[Polat and Güneş, 2007] Polat, K. and Güneş, S. (2007). Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. Applied Mathematics and Computation, 187(2):1017 – 1026.

[Quesenberry and Hurst, 1964] Quesenberry, C. P. and Hurst, D. C. (1964). Large sample simultaneous confidence intervals for multinomial proportions. Technometrics, 6(2):191–195.

[R Core Team, 2021] R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[Rafiei and Mendelzon, 1998] Rafiei, D. and Mendelzon, A. (1998). Efficient retrieval of similar time sequences using dft. Technical report, arXiv.

[Röhling et al., 2020] Röhling, S., Linne, A., Schellhorn, J., Hosseini, M., Dencker, T., and Morgenstern, B. (2020). The number of k-mer matches between two dna sequences as a function of k and applications to estimate phylogenetic distances. Plos One, 15(2):e0228070.

[Ruhaak et al., 2013] Ruhaak, L. R., Miyamoto, S., and Lebrilla, C. B. (2013). Developments in the identification of glycan biomarkers for the detection of cancer. Molecular & Cellular Proteomics, 12(4):846–855.

[Ruiz-Perez et al., 2020] Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., and Narasimhan, G. (2020). So you think you can pls-da? BMC Bioinformatics, 21(1):2.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4):406–425.

[Salcedo and Baldasano Recio, 1984] Salcedo, A. C. and Baldasano Recio, J. M. (1984). Fourier analysis of meteorological data to obtain a typical annual time function. Solar Energy, 32(4):479 – 488.

[Seeling et al., 2017] Seeling, M., Brückner, C., and Nimmerjahn, F. (2017). Differential antibody glycosylation in autoimmunity: sweet biomarker or modulator of disease activity? Nature Reviews Rheumatology, 13(10):621–630.

[Shanks, 1969] Shanks, J. L. (1969). Computation of the fast walsh-Fourier transform. IEEE Transactions on Computers, 100(5):457–459.

[Singleton, 1969] Singleton, R. (1969). An algorithm for computing the mixed radix fast Fourier transform. IEEE Transactions on Audio and Electroacoustics, 17(2):93–103.

[Slatko et al., 2018] Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next-generation sequencing technologies. Current protocols in molecular biology, 122(1):e59–e59. 29851291[pmid].

[Song et al., 2020] Song, Y., Cong, Y., Wang, B., and Zhang, N. (2020). Applications of Fourier transform infrared spectroscopy to pharmaceutical preparations. Expert Opinion on Drug Delivery, 17(4):551–571. PMID: 32116058.

[Stevens, 2007] Stevens, J. P. (2007). Intermediate Statistics: A Modern Approach. Routledge, 3rd edition.

[Stockwell et al., 2014] Stockwell, P. A., Chatterjee, A., Rodger, E. J., and Morison, I. M. (2014). Dmap: differential methylation analysis package for rrbs and wgbs data. Bioinformatics, 30(13):1814–1822.

[Strimmer et al., 2007] Strimmer, K., laure Boulesteix, A., and laure Boulesteix, A. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. In Briefings in Bioinformatics, pages 32–44.

[Strimmer et al., 2009] Strimmer, K., von Haeseler, A., and Salemi, M. (2009). Genetic distances and nucleotide substitution models. The Phylogenetic Handbook, pages 111–141.

[Tamura, 1992] Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+ c-content biases. Mol Biol Evol, 9(4):678–687.

[Tamura and Nei, 1993] Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. Molecular biology and evolution, 10(3):512–526.

[Tate and Bird, 1993] Tate, P. H. and Bird, A. P. (1993). Effects of dna methylation on dna-binding proteins and gene expression. Current opinion in genetics & development, 3(2):226–231.

[Thornton, 2019] Thornton, M. (2019). The invariance of spectral-kolmogorov-type statistics for estimating genomic similarity. In 2019 IEEE 49th International Symposium on Multiple-Valued Logic (ISMVL), pages 73–78. IEEE.

[Thornton, 2021] Thornton, M. (2021). Genomic DFT in R.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

[Tibshirani, 1997] Tibshirani, R. (1997). The lasso method for variable selection in the cox model. Statistics in Medicine, 16(4):385–395.

[Tomikawa et al., 2012] Tomikawa, J., Uenoyama, Y., Ozawa, M., Fukanuma, T., Takase, K., Goto, T., Abe, H., Ieda, N., Minabe, S., Deura, C., et al. (2012). Epigenetic regulation of kiss1 gene expression mediating estrogen-positive feedback action in the mouse brain. Proceedings of the National Academy of Sciences, 109(20):E1294–E1301.

[Tonegawa, 1983] Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature, 302(5909):575–581.

[Toyota et al., 1999] Toyota, M., Ahuja, N., Suzuki, H., Itoh, F., Ohe-Toyota, M., Imai, K., Baylin, S. B., and Issa, J.-P. J. (1999). Aberrant methylation in gastric cancer associated with the cpg island methylator phenotype. Cancer research, 59(21):5438–5442.

[Turin, 1960] Turin, G. (1960). An introduction to matched filters. IRE transactions on Information theory, 6(3):311–329.

[Uh et al., 2020] Uh, H.-W., Klarić, L., Ugrina, I., Lauc, G., Smilde, A. K., and Houwing-Duistermaat, J. J. (2020). Choosing proper normalization is essential for discovery of sparse glycan biomarkers. Molecular omics, 16(3):231–242.

[Van den Boogaart and Tolosana-Delgado, 2013] Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). Analyzing compositional data with R, volume 122. Springer.

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. Journal of Machine Learning Research, 9(Nov):2579–2605.

[Voss, 1992a] Voss, R. (1992a). Evolution of long-range fractal correlation and 1/f noise in dna base sequences. Physical Review Letters, 68:3805–3808.

[Voss, 1992b] Voss, R. F. (1992b). Evolution of long-range fractal correlations and 1/f noise in dna base sequences. Physical review letters, 68(25):3805.

[Waage and Guldberg, 1864] Waage, P. and Guldberg, C. (1864). Studier over affiniteten. Forhandlinger i Videnskabs-selskabet i Christiania, 1:35–45.

[Walley, 1996] Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):3–34.

[Wallingford and Ewing, 1989] Wallingford, R. and Ewing, A. (1989). Capillary electrophoresis. Advances in chromatography, 29:1–76.

[Walsh and Xu, 2006] Walsh, C. and Xu, G. (2006). Cytosine methylation and dna repair. DNA Methylation: Basic Mechanisms, pages 283–315.

[Wang et al., 2010] Wang, H., Meng, J., and Tenenhaus, M. (2010). Regression modelling analysis on compositional data. In Handbook of Partial Least Squares, pages 381–406. Springer.

[Wang et al., 2013] Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery, 26(2):275–309.

[Wasserman, 2006] Wasserman, L. (2006). All of Nonparametric Statistics. Springer Texts in Statistics. Springer New York Inc.

[Watanabe and Maekawa, 2010] Watanabe, Y. and Maekawa, M. (2010). Methylation of dna in cancer. Advances in clinical chemistry, 52:145–167.

[Wen et al., 2014] Wen, J., Chan, R. H., Yau, S.-C., He, R. L., and Yau, S. S. (2014). K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene, 546(1):25–34.

[Wilks, 1932] Wilks, S. S. (1932). Certain generalizations in the analysis of variance. Biometrika, pages 471–494.

[Wold, 1966] Wold, H. (1966). Estimation of principal components and related models by iterative least squares. Multivariate Analysis, pages 391–420.

[Woodward, 2014] Woodward, P. M. (2014). Probability and information theory, with applications to radar: international series of monographs on electronics and instrumentation, volume 3. Elsevier.

[Wu et al., 2015] Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P., and Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic acids research, 43(21):e141–e141.

[Wu et al., 2000] Wu, Y., Agrawal, D., and Abbadi, A. E. (2000). A comparison of dft and dwt based similarity search in time series databases. In Proceedings of the Ninth International Conference on Information and Knowledge Management, pages 488–495. ACM.

[Xie et al., 2019] Xie, C., Leung, Y.-K., Chen, A., Long, D.-X., Hoyo, C., and Ho, S.-M. (2019). Differential methylation values in differential methylation analysis. Bioinformatics, 35(7):1094–1097.

[Yao et al., 2019] Yao, Q., Chen, Y., and Zhou, X. (2019). The roles of microrna s in epigenetic regulation. Current opinion in chemical biology, 51:11–17.

[Yin, 2020] Yin, C. (2020). Phylogenetic analysis of DNA sequences or genomes by Fourier transform.

[Yin et al., 2014] Yin, C., Chen, Y., and Yau, S. (2014). A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. Journal of Theoretical Biology, 359:18–28.

[Yin and Yau, 2015] Yin, C. and Yau, S. S.-T. (2015). An improved model for whole genome phylogenetic analysis by Fourier transform. Journal of Theoretical Biology, 382:99 – 110.

[Yin et al., 2017] Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on dna binding specificities of human transcription factors. Science, 356(6337).

[Zhang et al., 1988] Zhang, W. et al. (1988). Shift-invariant pattern recognition neural network and its optical architecture. In Proceedings of annual conference of the Japan Society of Applied Physics.

[Zhang et al., 2020] Zhang, Y., Park, C., Bennett, C., Thornton, M., and Kim, D. (2020). Hisat-3n: a rapid and accurate three-nucleotide sequence aligner. Genome Research.

[Zheng-Bradley et al., 2017] Zheng-Bradley, X., Streeter, I., Fairley, S., Richardson, D., Clarke, L., Flicek, P., and Consortium, . G. P. (2017). Alignment of 1000 genomes project reads to reference assembly grch38. Gigascience, 6(7):gix038.

[Zhou et al., 2007] Zhou, Y., Zhou, L.-Q., Yu, Z.-G., and Anh, V. (2007). Distinguish coding and noncoding sequences in a complete genome using Fourier transform. In Third International Conference on Natural Computation (ICNC 2007), volume 2, pages 295–299. IEEE.