Statistical Topology of Reticulate Evolution

Kevin Joseph Emmett

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Statistical Topology of Reticulate Evolution

Kevin Joseph Emmett

Reticulate modes of genetic exchange can confound traditional methods of establishing evolutionary relationships among sets of related organisms. Increasing data has pointed to the prevelance of these modes, and underscored the lack of a unified mathematical approach to capturing and representing the scale and frequency of these events. This thesis contains result of applying new mathematical methods drawn from applied and computational topology to the problem of measuring reticulate evolution in molecular sequence data. In so doing, new techniques are established for constructing topological representations and extracting statistical patterns from biological data sets. We apply our approaches to several types of molecular sequence data, include bacteriophage, influenza, and pathogenic bacteria. We also consider patterns of intranuclear chromatin folding in bacteria and humans.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Darwin's *On the Origin of Species* contains a single figure, depicting the ancestry of species as a branching genealogical tree, or *phylogeny* [38] (see Figure 1.1). Darwin argued that evolution was mediated by descent with modification; that is, the gradual change in heritable traits under the pressure of natural selection. Since that time, the tree structure has been the dominant framework to understand, visualize, and communicate discoveries about evolution. Indeed, an important aim of evolutionary biology has been expanding the *universal tree of life*, the set of evolutionary relationships among all extant and extinct organisms on Earth [18].

Traditionally, evolutionary relationships were established on the basis of phenotype, i.e. the observable traits of each organism. With the advent of molecular models of evolution and rapidly increasing genomic sequence data, the genotype has supplanted phenotype as the primary focus of evolutionary studies. Molecular phylogenetics has become established as the standard tool for inferring phylogenetic relationships. However, a tree is accurate only if the Darwinian model of descent with modification is the sole process driving evolution. It has long been recognized that there exist alternative evolutionary processes that can allow organisms to exchange genetic material through means beyond reproduction [5]. Notable examples include horizontal gene transfer in bacteria, species hybridization in plants, and

meiotic recombination in eukaryotes. Collectively, these processes are known as *reticulate evolution*. These stand in contrast to descent with modification, an example of *clonal evolution*.[1] Increasing genomic data, powered by new high-throughput sequencing technologies, has shown that these reticulate processes are more prevalent than originally expected. For some, this has called into question the tree of life hypothesis as an organizing principle and prompted the search for new ways of representing evolutionary relationships [43, 112].

This thesis presents a new approach to quantifying and representing reticulate evolutionary processes using recently developed ideas from algebraic and computational topology. The methods we employ fall under the collective heading of *topological data analysis* (henceforth TDA), a new branch of applied topology concerned with inferring structure in high-dimensional data sets. The thesis consists of three aims: (1) introduce the methods of TDA and their application to biological and genomic data; (2) develop approaches tailored to the unique features of molecular sequence data; and (3) apply these approaches to a wide range of biological problems in which reticulate processes are believed to play an important role.

In the following brief introduction, we survey salient aspects of molecular evolution, the tree paradigm, and the challenges posed by reticulate processes. We then introduce the idea of representing evolution as a topological space and give a flavor of the results to be discussed.

1.1 Molecular Evolution and the Tree Paradigm

The combination of Darwin's theory of natural selection with Mendelian genetics led to the *modern evolutionary synthesis*, outlined in the first half of the twentieth century in pioneering works by Ronald Fisher, Sewall Wright, JBS Haldane, and others.[2] The modern

[1] Clonal and reticulate evolution are also known by the terms *vertical* and *horizontal* evolution, respectively.

[2] See [75] and [67] for comprehensive historical detail.

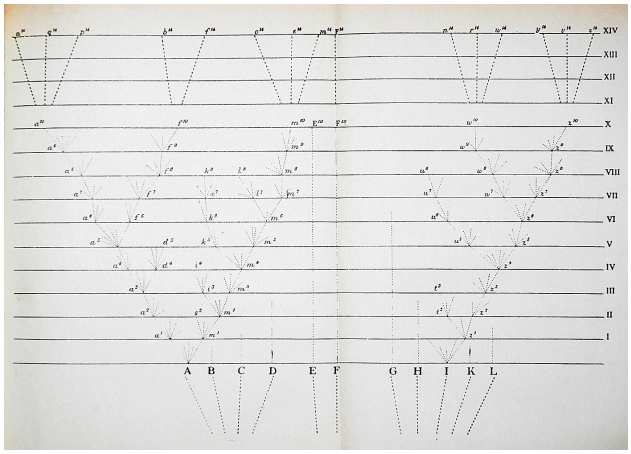


Figure 1.1: The only figure in Darwin's *On the Origin of Species*. Darwin argued for descent with modification and natural selection as the driving processes underscoring evolution. In this figure, Darwin sketched his idea for how diverging species would result in a tree structure. Reproduced from [38].

synthesis was based largely on an analysis of distributions of allele frequencies in distinct populations, the purview of classical population genetics. The field was placed on a molecular foundation with Watson and Crick's discovery of the DNA double-helix in 1953 [139]. These developments led to the establishment of *molecular evolution*, the analysis of how processes such as mutation, drift, and recombination act to induce changes in populations and species.

The information underlying an organism's form and function is encoded in its genome, the complete sequence of DNA contained in each cell. The genome can be represented as a string of nucleotides, indexed by position. Embedded within the genome are regions defining the genes which code for functional proteins, as well as non-coding regions which have as-yet unknown function.[3] When an organism reproduces, either sexually or asexually, a complete copy of the genomic information is passed to the offspring. Because the molecular mechanisms that control this copying are not exact, errors in replication are introduced. These errors can take the form of single point mutations (or single nucleotide polymorphisms, SNPs), small insertions and deletions of a few nucleotides (indels), or larger effects including

[3]In humans, only 1.5% of the genome is protein-coding, the rest largely non-functional. [88]. Up to 5-8% of the human genome is believed to consist of endogenous retroviruses, dead viruses which have integrated their genome into the human genome.

copy numbers variations (CNVs) and chromosomal duplications.[4] Under the neutral theory of evolution, the majority of these errors will have very little impact, either positive or negative, on the descendant organism. A small fraction of mutations will result in an appreciable fitness differential compared to other organisms, and it is on these organisms that natural selection will act.

While molecular biology has largely focused on the biochemical and biophysical mechanisms underlying these processes, *molecular phylogenetics* has focused on the comparative analysis of macromolecular sequences to infer genealogical and evolutionary relationships. Molecular phylogenetics began with Emile Zuckerkandl and Linus Pauling's recognition in the early 1960's that the information encoded in a set of molecular sequences could itself be used as a document of evolutionary history [150, 151]. It became clear that given two sequenced organisms, counting the differences between their respective sequences could be used as a quantitative measure of the amount of evolutionary divergence between the two organisms. If one had a larger set of sequenced organisms, computing the complete set of pairwise distances yields a *distance matrix*. From the distance matrix, one attempts to associate a tree to the data such that pairwise distances along the tree are close to the observed pairwise sequence distances. Walter Fitch and Emanuel Margolish helped popularize this approach by constructing a weighted least squares approach to fitting phylogenetic trees from distance matrices [59]. Since that time, the development of numerical approaches for inferring evolutionary relationships has evolved into a mature discipline and the use of molecular sequence data to infer phylogeny has become a standard practice across a wide range of biology and ecology. While other approaches to tree inference have been developed, including parsimony, maximum likelihood (ML) and Bayesian methods, our focus will be on distance matrix methods because of their close connection with the topological ideas we employ later.

[4]Mutation rates vary across species. In humans, 10^{-8} per site per generation. In single cell bacteria, 10^{-10} per site per generation.

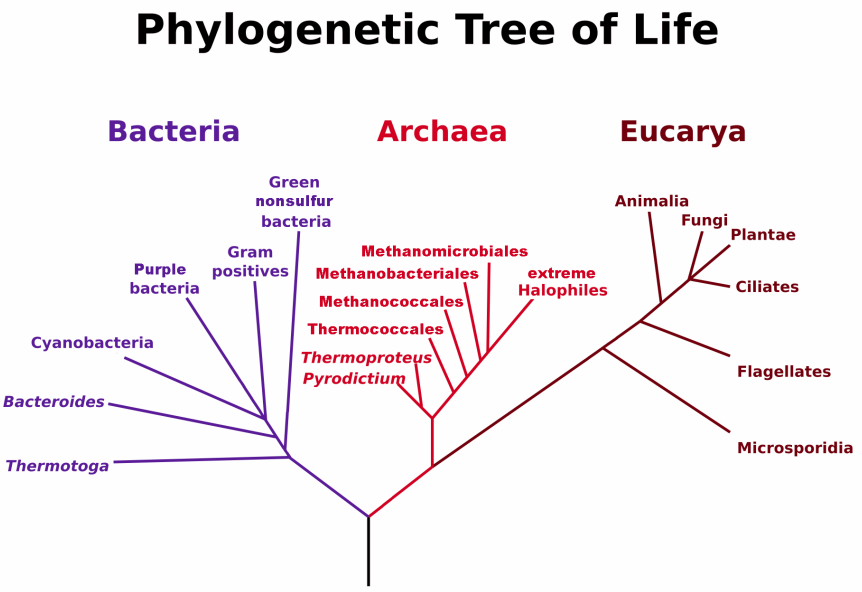


Figure 1.2: Carl Woese's three kingdom tree of life. Using 16S subunit ribosomal RNA, Carl Woese identified archaea as a distinct phylogenetic kingdom. Previously, based on morphological similarity (specifically, unicellular and lacking a nucleus), archaea had been grouped with bacteria. This result was an early success for molecular phylogenetics and the use of conserved gene segments for molecular classification. Figure adapted from [145].

One important early result from molecular phylogenetics was Carl Woese's organization of bacteria, eukarya, and archaea into the three domains of life [144]. Prior to Woese, there were two recognized domains of life: prokaryotes, single-celled organisms lacking a nucleus, and eukaryotes, multi-celled organisms with an enveloped nucleus. Using 16S subunit ribosomal RNA sequencing, Woese discovered that the prokaryotic domain actually split into two evolutionarily distinct groups. One of these, which he termed *archaebacteria* was more closely related to eukaryotes than were there the rest of the prokaryotes. This led to the three-domain system of life (see Figure 1.2).

This work had several important consequences. First, it established the use of molecular data to inform about large-scale patterns of evolutionary history. Using only morphological data had led to an inconsistent classification of archaea. Second, it positioned 16S rRNA profiling as the primary source of data for use in comparative genomics. The use of this genomic region was justified on the basis of being one of the few universal gene segments

that is conserved across all species. Constructing a universal tree is predicated on there being orthologous genes, shared genes related through speciation events, that can provide a common foundation for comparative study. Finally, it solidified the tree paradigm as an organizing principle for relating extant species. Even though reticulate processes had been known since the early twentieth century[5], the idea that evolutionary relationships should be described by a bifurcating tree had been paramount since Darwin. Reticulate processes were either ignored completely, or expected to occur at such low frequencies that they need not be considered.

1.2 Reticulate Processes and the Universal Tree

Despite the significant impact of Woese's observation, there remains a subtle difficulty, which Woese himself would come to contemplate in later work [143, 66]. Woese's phylogeny was based on only 1,500 nucleotides in the ribosomal RNA, less than 1% of the total length of a typical bacterial genome (see [37]). Even more striking to consider, this accounts for less than 0.00005% of the human genome. While recent work has developed approaches for constructing reference trees from larger gene sets [34], the fact remains that the vast majority of genomic information is *not* incorporated into the tree.

The reason for this situation is twofold. First, not all genes are shared universally across all species. In constructing a phylogenetic tree using sequence data, only genes that are present across all species are informative. Second, even among universal genes, the presence of reticulate evolutionary processes will confound systematic analysis. The model of a bifurcating tree will be consistent only if all loci share the same pattern of bifurcation. When organisms can exchange genetic material by means other than clonal reproduction, the ancestral relationships between species will depend on which genomic regions are used. If one were to use two different genomic regions, two different tree topologies may be gener-

[5]Beginning with Frederick Griffith's experiments in 1928 showing that non-virulent strains of *Streptococcus pneumoniae* could acquire virulence factors by being exposed to dead virulent strains.

ated, with conflicting phylogenetic relationships. It remains an open question how to best construct a consistent evolutionary history from conflicting phylogenetic signals[6]

Historically, reticulate processes were believed to occur at such a low frequency that they could be safely ignored when considering evolutionary relationships. However, new genomic data has shown that, particularly in microorganisms such as bacteria and archaea, reticulate processes are much more prevalent than originally expected [111]. Incompatibilities in the tree paradigm now appear as the rule, not the exception, which has led to calls for new representations of evolutionary relationships [43, 44]. Many have argued that, in light of new genomic evidence, the very notion of a universal tree of life must be discarded [82, 83]. This point has been argued most strongly by Ford Doolittle. In Figure 1.3, we see Doolittle's simplified representation of Tree of Life as it stands today, including only the most well-known examples of reticulations, the acquisition of mitochondria and chloroplasts from bacterial ancestors. We also see his representation of the Tree of Life as it would stand if additional large-scale reticulations were contained – no long is it clear that the tree is an appropriate metaphor.

Finally, reticulate evolutionary processes are of more than just theoretical concern. In HIV, frequent homologous recombination confounds our understanding of the epidemic's early and present history [23]. In influenza, segmental gene reassortments lead to antigenic novelty and the emergence of epidemics [106]. In several pathogenic bacteria, including *E. coli* and *S. aureus*, horizontal gene transfer has been responsible for the spread of antibiotic resistance genes [2, 39]. For example, the 2011 German *E. coli* outbreak was caused by a strain of *E. coli* that had acquired the ability to produce Shiga toxin [118].

[6]There exists a cottage industry of methods for aggregating conflicting *gene trees* into a consensus *species tree*, see [99].

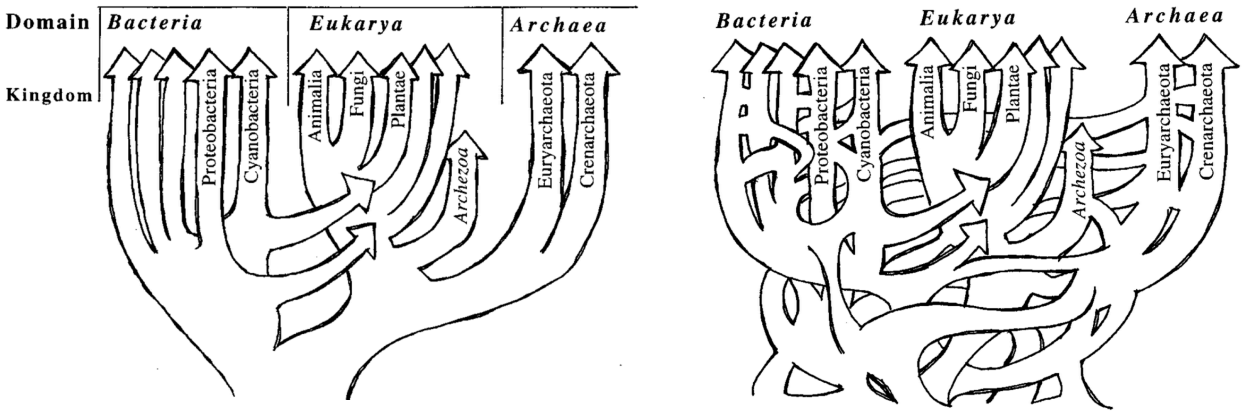


Figure 1.3: Left: W Ford Doolittle's representation of the consensus universal tree of life. Only the most well known reticulations are reflected: the endosymbiosis of mitochondria and chloroplasts. Right: Doolittle's representation of the universal tree of life with reticulate evolution. While the three domains of life are still recognizable, patterns of divergence no longer follow a strictly treelike model. (From *Science*, vol. 284, issue 5423, page 2127. Reprinted with permission from AAAS.)

1.3 Evolution as a Topological Space

We propose the use of new computational techniques, borrowed from the field of applied topology, to capture and represent complex patterns of reticulate evolution.

Topology as a mathematical field is concerned with properties of spaces that are invariant under continuous deformation. Such properties can include, for example, connectedness and the presence of holes. Two objects are considered topologically equivalent if they can be deformed into one another without introducing any cuts or tears. As a paradigmatic example, consider the coffee mug and the donut (Figure 1.4). While seemingly different, it is not difficult to see that both objects consist of a single connected component that is wrapped around a single hole. Were the objects smoothly pliable they could be freely deformed into one another. Topologically, the two objects are equivalent.[7]

[7]The two objects are topologically equivalent to a solid torus, which is represented as $D^2 \times S^1$, a solid two-dimensional disk wrapping around a circle.

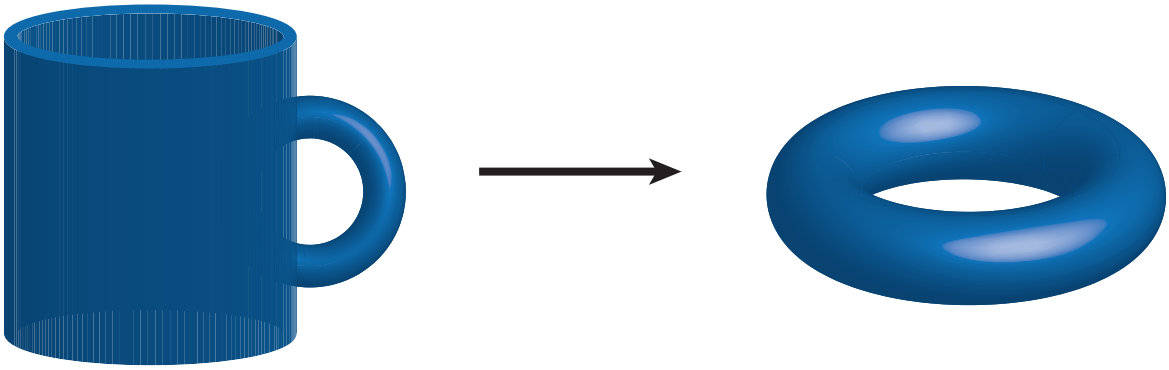


Figure 1.4: The paradigmatic example of topological equivalence. The coffee mug can be continuously deformed into the donut and are therefore topologically equivalent. Both exhibit the topology of a solid torus ($D^2 \times S^1$).

Algebraic topology quantifies our intuitive notions of shape by associating algebraic structures to different invariants. For our purposes, the most relevant invariants will be the *Betti numbers*. We give a more complete characterization of Betti numbers in Chapter 2, but the intuition is as follows. The Betti numbers are a collection of integers indexed by an integer n describing the connectivity of a space at different dimensions. First, we can think of b_0 as representing the number of connected components, or clusters, in our space. Next, we can think of b_1 as representing the number of one-dimensional loops in our space. Equivalently, this is the number of cuts needed to transform the space into something simply connected. Higher Betti numbers, b_n for $n > 1$ will correspond to higher dimensional holes. In our coffee mug example, because both objects have the same Betti numbers ($b_0 = 1$, $b_1 = 1$, and $b_n = 0$ for $n > 1$), they can be considered topologically equivalent. Our goal in this work will be to adopt a similar perspective as this example and characterize evolutionary spaces as topological spaces using their Betti numbers.

To give the very simplest example, consider Figure 1.5. The example presents two possible scenarios describing the evolutionary relationships of three species, labeled a, b, and c. The

objects are to be read such that moving vertically corresponds to moving backwards in time. Branch lengths will correspond to some notion of evolutionary divergence. Internal vertices represent extinct ancestors of the three species, up to the root of the tree, r, which represents the most recent common ancestor. On the left, we have a simple tree topology relating the three species. Considering the shape of the tree, there is a single connected component, giving $b_0 = 1$. Further, we see that there are no loops formed by the branches, giving $b_1 = 0$. The object is trivially contractible, a property which will hold for all tree topologies. On the right, we have a reticulate topology relating the three species. We can envision species b as being the reticulate offspring of parents ancestral to species a and c. That is, species b carries unique genetic material from both species a and species c. To account for this, two branches merge into the vertex that is directly ancestral to b. Considering the shape, there is again a single connected component, giving $b_0 = 1$. However, because of the reticulate event mixing material from a and c, there is now a loop formed in the topology, giving $b_1 = 1$ The object is no longer treelike and is characterized by a nontrivial topology. The Betti numbers capture the essential difference in the two evolutionary histories.

Consider again Darwin's branching phylogeny (Figure 1.1) and Doolittle's modified tree accounting for reticulate evolution (Figure 1.3). The two objects can be imagined to be representations of two different topological spaces. Darwin's branching phylogeny is a tree and hence trivially contractible ($b_n = 0$ for $n > 0$). In contrast, Doolittle's construction has a much more complex topology, with loops being formed at points where reticulate events have occurred. The object will have nonvanishing Betti numbers, which will be associated with the amount of reticulation. The remainder of this thesis focuses on expanding this idea and applying it to real data sets with the goal of measuring the prevalence and scale of reticulate evolutionary events. Our aim will be to characterize reticulate exchanges of genetic material by the parental sequences involved in the exchange, by the amount and identity of material exchanged (i.e., the genes or loci involved), and the frequency with which similar exchanges occur. Several important questions will be dealt with, such as how

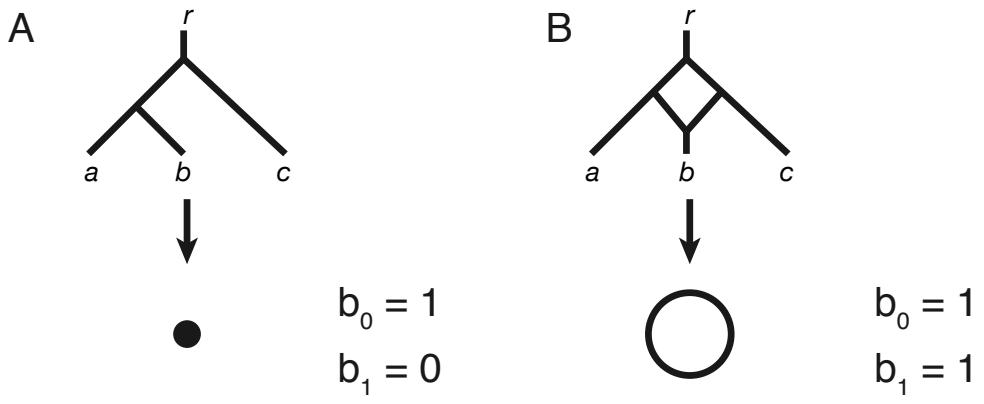


Figure 1.5: (A) A simple treelike phylogeny is contractible to a point. (B) A reticulate phylogeny that is equivalent to a circle and not contractible without a cut. The two spaces are not topologically equivalent and can be distinguished by their Betti numbers.

to construct topological spaces from finite samples, how to make comparisons among gene sets, and how to make statistical statements about reticulate events. We will address these questions, and in doing so develop new techniques to construct and extract topological and statistical information from evolutionary data. In doing so, we provide a fuller understanding of evolutionary relationships than allowed by current phylogenetic methods.

1.4 Thesis Organization

The remainder of this thesis is organized as follows.

In Chapter 2 we present background material on the topics discussed in this thesis. This discussion is chiefly structured into two pieces: (1) background on phylogenetics and population genetics, and (2) background on the methods we use from TDA.

In Part I, we develop two complementary approaches for analyzing sequence data using TDA. In Chapter 3, we propose methods of constructing topological spaces that generalize standard constructions but are suite to the particular requirements of phylogenetic applications. We draw on previous work in phylogenetic networks and use homology to provide a quantitative assessment of reticulate processes. This work was published in [53]. In Chapter

4, we develop methods for performing statistical inference using summary statistics computed using methods from TDA. This is the first such use of TDA as a tool for performing parametric inference and should generalize to a wide range of application settings. This work was published in [54]

In Part II, we apply our approach to several problems in evolution and genomics. In Chapter 5 we study bacteriophages. In Chapter 6 we study influenza. In Chapter 7 we study pathogenic bacteria and use topological techniques to represent the spread of antibiotic resistance. In Chapter 8 we use population data to measure human recombination rates and identify recombination hotspots. We identify variation in recombination hotspots in different human populations. In Chapter 9 we analyze Hi-C data to explore patterns of chromatin folding in the nucleus in both prokaryotic and human datasets.

Finally, in Chapter 10 we summarize these results and present future research directions.

Chapter 2

Background

This thesis uses newly developed approaches from applied topology to study problems in evolutionary biology and genomics. In this chapter we provide background material to motivate our approach. In Section 2.1 we introduce the models of evolution and types of genomic data we will consider. In Section 2.2 we provide a self-contained introduction to the primary methods of topological data analysis, including persistent homology and Mapper. Finally, in Section 2.3 we give simple examples of how the tools from TDA can be informative about reticulate evolution.

2.1 Biology

In this section we present a basic introduction to molecular sequence data: what the data looks like, the processes by which it is generated, and the methods by which it is analyzed. Particular attention is paid to modes of reticulate evolution. Exposition for specific applications can be found in their respective individual chapters.

2.1.1 Genes and Genomes

The information required to express an organism's biological form and function is contained in the genome. At least one copy of the genome is packaged inside each cell of an organism. Physically, the genome is manifest as a polymer chain of nucleic acid, built on an alphabet of four nucleotide monomers: adenine, cytosine, guanine, and thymine. Abstractly, the genome is represented as a linear sequence of characters defined over the alphabet $\{A, C, G, T\}$.[1] Contained in this sequence are subsequences representing genes, which code for the protein products that ultimately affect function. Further embedded in the genome is a complex regulatory pattern of transcription factors controlling the expression of particular genes and directing cellular differentiation and development.

Following the central dogma of biology, DNA is transcribed into RNA, RNA is translated into amino acids, and amino acids are folded into proteins [36]. Proteins comprise the functional unit of biology.

Beyond simply coding for function, the genome includes an imprint of the evolutionary history that gave rise to the organism. By comparing the genomes of multiple organisms, inferences can be drawn about the evolutionary relationships among extant organisms as well as the processes that generated observed diversity. The field concerned with exploring these relationships is *comparative genomics*.

2.1.2 Evolutionary Processes

Evolution describes the gradual change in phenotypes arising from random variation and subject to natural selection. The processes giving rise to diversity can be classified into two types: clonal and reticulate.

[1] The linear representation can be misleading, as many organisms, primarily viruses and bacteria, have circular genomes.

2.1.2.1 Clonal Evolution

Clonal evolution, or vertical evolution, is a process of self-reproduction whereby genetic material is transferred directly from parent to offspring. Population diversity is generated by stochastic mutation and maintained over multiple generations by random drift.

It is clonal evolution that Darwin had in mind when he described the idea of descent with modification, whereby a parent passes genomic information to an offspring subject to random drift. Importantly, because there is always a direct parent–offspring relationship, clonal evolution can be modeled with a binary tree model.

2.1.2.2 Reticulate Evolution

Reticulate evolution, or horizontal evolution, refers to exchange or acquisition of genetic material via processes that do not reflect a direct parent–offspring relationship. As we will see, these processes can make inferences about historical evolutionary relationships difficult. Different types reticulate processes occur in different types of organisms (summarized in Table 2.1).

Viruses replicate by infecting a host cell and then using the host cell machinery and resources to produce multiple copies of viral genetic material. The genetic material is then packaged into new virus particles which are shed off in order to infect new cells. Reticulation can occur when two virus particles coinfect the same host cell. During the replication process, genetic material can be exchanged in one of two ways: *reassortment* or *recombination* (the two processes are contrasted in Figure 2.1). Reassortment occurs in viruses whose genomes are segmented, such as influenza. Segments are similar to chromosomes, such that a single virus particle will contain a single copy of each segment. Coinfection of a single cell with two independent viruses results in packaging of segments taken from different virus particles. The result viral progeny will then be a genetic mixture of segments from each parental strain. Recombination, more common in non-segmented viruses such as HIV, involves a break-rejoin mechanism during the replication process. Here, an error in the polymerase during

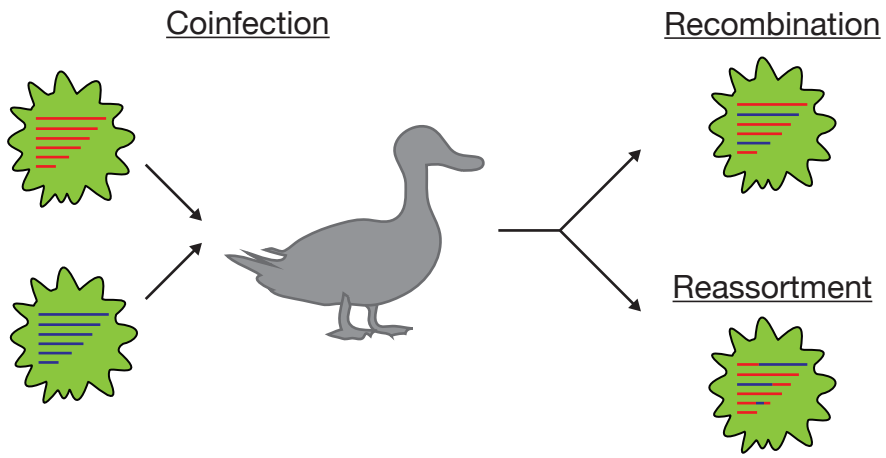


Figure 2.1: The two modes of viral reticulation. Coinfection of the same host cell can lead to either reassortment, in which whole viral segments are exchanged, or recombination, in which breakpoints can occur within segments. The former process is common in influenza, the latter in HIV. The end result, however, is a novel virus particle which shares genetic information from both parents.

replication can result in an incomplete copy of the genome (a break). At this point, several cellular processes involved in repair can be recruited to complete the replication process using a homologous region. If coinfection has occurred, it is possible for these processes to initiate repair using material from a different parental strain. The outcome will be novel genetic material that includes a crossover from one strain to another. Break-rejoin crossover is a type of *homologous recombination*.

In bacteria and other prokaryotes, reticulate evolution can occur when foreign DNA from a donor is acquired by a target organism and integrated into its genome. Three generic mechanisms have been identified, depending on the route by which foreign DNA is acquired [111]:

1. *Conjugation*. Direct cell-to-cell contact between donor and recipient resulting in transfer of plasmid.

2. *Transformation*. Foreign DNA acquired via uptake from freely circulating DNA in the environment.

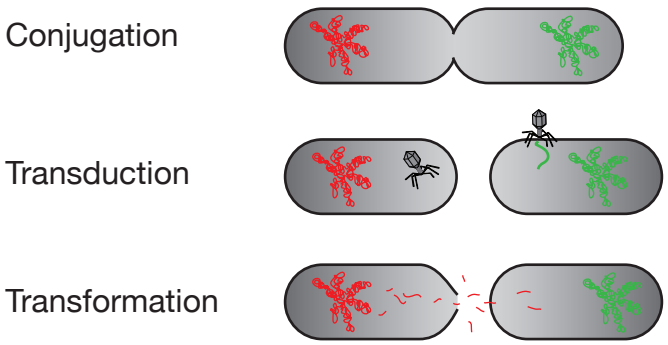


Figure 2.2: Three modes of viral reticulation. (1) Conjugation, in which direct cell-to-cell contact results in transfer of genetic material; (2) Transformation, in which foreign DNA is acquired via uptake from freely circulating DNA in the environment; and (3) Transduction, in which exchange of genetic material is mediated by a virus or phage particle.

3. *Transduction.* Virus-mediated transfer for foreign DNA from an infected donor cell.

A visualization of these three mechanisms is shown in Figure 2.2. Because these mechanisms can often lead to the acquisition of novel sequences coding for genes not in the recipient organism, reticulate evolution in prokaryotes is often called *horizontal gene transfer* or *lateral gene transfer*.

In eukaryotes, several reticulate processes have been identified. We mention two such processes: hybrid speciation and meiotic recombination. These two processes act at very different scales, however the outcome is the same: a unique offspring with genetic material drawing from both parents.

First, hybrid speciation refers to the cross-breeding of animals or plants of different species. This mixing of genetic material can lead to the development of a third species with a phenotype distinct from both parents. Hybrid speciation was originally believed to be a rare occurrence in nature and hybrid offspring to be infertile. However, recent genomic data has demonstrated that hybrid speciation occurs quite frequently in plants [4, 5]. Indeed, Mendel's early experiments in hybridization were themselves an artificially induced form of reticulate evolution.

Second, meiotic recombination refers to a specialized process for generating diversity

that occurs in sexually-reproducing polyploid organisms, such as humans, during meiosis. Meiosis is the process by which a single cell containing n copies of each chromosome results in four distinct cells each with $n/2$ copies of each chromosome. These special cells are called gametes. Sexual reproduction consists of the fusion of two gametes during fertilization to form a zygote, which ultimately develops into a viable offspring. Meiosis is a multi-step process consisting of an initial round of DNA replication followed by two rounds of cell division. Meiotic recombination occurs after the initial round of DNA replication and prior to cell division. After DNA replication, there are two copies of each homologous chromosome that are joined at a centromere. The two sets of chromosomes then pair with each other and exchange DNA through physical interactions known as crossovers.[2] This is another example of homologous recombination and results in new allelic patterns mixing genetic information from both parents.[3] After crossover occurs, two phases of cellular division result in gametes with $n/2$ copies of each chromosome.

At this point, one might wonder about sexual reproduction – an offspring can be seen as a hybridization of genetic material donated from both mother and father. On the one hand, the answer could be yes, particularly because the presence of meiotic recombination involves a shuffling of genomic material such that the chromosome each parent donates is a unique combination of alleles not previously present in the donor organism. On the other hand, the answer could be no, because both mother and father donate a complete copy of the genome to the offspring. Each copy can be considered as an independent transfer of genetic information defining both a matrilineal and a patrilineal line of inheritance. Indeed, researchers in human population genetics generally distinguish between these two cases – looking at genomic regions that do not recombine they define a matrilineal common ancestor known as *Mitochondrial Eve* and a patrilineal common ancestor known as *Y-chromosomal Adam*. Mathematically, the evolutionary relationships of a population of N organisms with

[2]These crossovers have been shown to not occur randomly, but rather at recombination hotspots regulated by binding motifs for by the PRDM9 protein [11, 24].

[3]Patterns of shared alleles define the concept of *linkage*.

Table 2.1: Reticulate processes in biology across kingdoms

Organism	Process	Description
Virus	Reassortment	Exchange of discrete genomic segments
	Recombination	Intragenomic homologous crossover
Bacteria	Transformation	Acquisition of foreign DNA in environment
	Transduction	Viral-mediated exchange
	Conjugation	Cell-to-cell contact and exchange
Eukaryotes	Meiotic Recombination	Homologous crossover during meiosis
	Hybrid Speciation	Fertilization across species boundaries

ploidy n are often analyzed as a haploid population of size nN with random mating.

The presence of reticulate processes in a set of organisms can be most clearly identified by comparing phylogenetic relationships built from different genomic segments. A general practice is to construct the set of *gene trees* which reflect ancestral branching patterns at specific loci. If a reticulate event has occurred, it implies that the branching patterns of different genes will not agree. A subfield of comparative genomics is concerned with building *species trees* from sets of gene trees [99].

However, in the case where there is substantial disagreement among gene trees, the very notion of a species tree may be flawed. Traditionally, evolutionary biology has concerned itself with characterizing relationships in light of vertical evolution alone. However, increasing evidence has pointed to the important role played by horizontal evolution, particularly in prokaryotic evolution [65, 64]. Between 10% to 16% of the *E. coli* genome is believed to have arisen from horizontal gene transfer [111].

2.1.3 Mathematical Models of Evolution

Mathematical population genetics is concerned with properties of populations as they are subject to evolutionary forces over long time scales. These forces include natural selection, genetic drift, mutation, and recombination. Historically the input data for population genetics models was comparative studies of allele frequencies across populations. These studies

have primarily been replaced by large-scale genomic surveys which have provided unprecedented insight into ancient population structure and historical migrations.

These models allow two things: (1) simulate genomic data under realistic processes and (2) build statistical models to estimate biological parameters from data.

2.1.3.1 The Wright-Fisher Model

The Wright-Fisher model is a forward time simulation of an evolving population. In the simplest case, the model describes neutral evolution of a constant population size with no structure and constant genome length. The model proceeds in units of generations. At each generation, a member of the population is an offspring of a randomly selected ancestor from the previous generation. This offspring inherits its ancestors genomes, with mutations introduced at some base rate μ. A member of previous generation with no offspring will be considered extinct.

2.1.3.2 The Coalescent Process

The coalescent process is a stochastic model that generates the genealogy of individuals sampled from an evolving population [136]. The genealogy is then used to simulate the genetic sequences of the sample. This model is essential to many methods commonly used in population genetics. Starting with a present-day sample of n individuals, each individual's lineage is traced backward in time, towards a mutual common ancestor. Two separate lineages collapse via a coalescence event, representing the sharing of an ancestor by the two lineages. The stochastic process ends when all lineages of all sampled individuals collapse into a single common ancestor. In this process, if the total (diploid) population size N is sufficiently large, then the expected time before a coalescence event, in units of $2N$ generations, is approximately exponentially distributed:

$$P(T_k = t) \approx \binom{k}{2} e^{-\binom{k}{2}t}, \tag{2.1}$$

where T_k is the time that it takes for k individual lineages to collapse into $k-1$ lineages.

After generating a genealogy, the genetic sequences of the sample can be simulated by placing mutations on the individual branches of the lineage. The number of mutations on each branch is Poisson-distributed with mean $\theta t/2$, where t is the branch length and θ is the population-scaled mutation rate. In this model, the average *genetic distance* between any two sampled individuals, defined by the number of mutations separating them, is θ.

The coalescent with recombination is an extension of this model that allows different genetic loci to have different genealogies. Looking backward in time, recombination is modeled as a splitting event, occurring at a rate determined by population-scaled recombination rate ρ, such that an individual has a different ancestor at different loci. Evolutionary histories are no longer represented by a tree, but rather by an *ancestral recombination graph*. Recombination is the component of the model generating nontrivial topology by introducing deviations from a contractible tree structure, and is the component which we would like to quantify. Coalescent simulations were performed using `ms` [72].

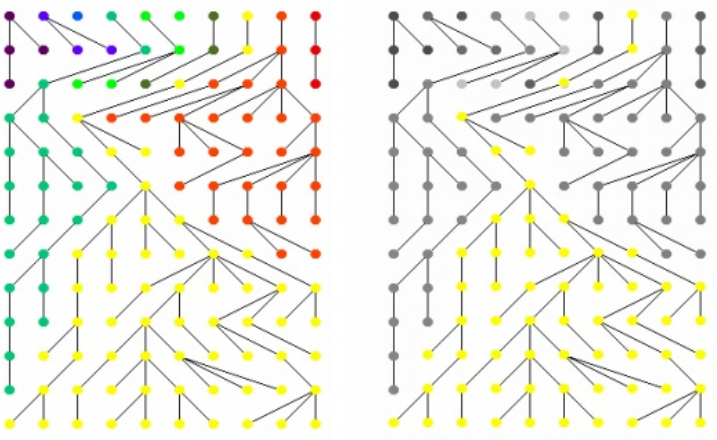
2.1.3.3 Metrics on Sequences

Evolutionary models require a notion of genetic divergence between sequences. This leads to a discussion of the types of metrics that can be put on sets of sequences.[4]

The simplest model, and the one most commonly adopted in this thesis, is the Hamming metric, which simply counts the proportion of sites that differ between two aligned sequences. For example, for two sequence $s_1 = ACTTGAC$ and $s_2 = AAGTGGC$, $d_H(s_1, s_2) = 3/7$. In general, the Hamming metric will underestimate divergences by not accounting for the possibility of back mutations.[5]

[4]Before sequences can be compared, they must first be *aligned*. A sequence alignment arranges the characters in a set of sequences into columns such that individual characters sharing an evolutionary identity are in the same column. Alignment is necessary because random insertions and deletions can change the relative . The difficulty of performing an alignment will largely depend on the amount of evolutionary divergence in the set of sequences under consideration. Sequence alignment is a well studied topic but largely beyond the scope of this thesis, where we assume sufficient sequence similarity such that alignment can be performed with high confidence.

[5]A double mutation of the form $A \to C \to A$.



(a) Wright-Fisher Model (b) Coalescent Model

Figure 2.3: Two models for simulating evolutionary data. On the left, the Wright-Fisher model simulates a sample of n individuals in the forward direction. At each generation, n offspring choose a parent from the previous generation at random. After t generations, some initial lineages will have died off, while others will become dominant in the population. On the right, the coalescent model simulates the sample in the reverse direction. At each reverse generation, individuals merge, or coalesce, with some probability, until they reach a single most recent common ancestor (MRCA). The intuition behind the approach is that lineages that have gone extinct will not contribute to the present day observed diversity, are therefore inaccesible, and do not need to be simulated. This approach reduces the data that needs to be simulated and increases the computational performance of the models.

More biologically motivated models will introduce corrections to account for assumptions about how sequences evolve. These assumptions include the base frequency of each nucleotide as well as the substitution rates for each type of mutation. The simplest of these models is the *Jukes-Cantor model*. This model defines an equal substitution rate μ. Inverting the probability of an alteration gives the divergence. The Jukes-Cantor metric is defined as

$$d_{JC} = -\frac{3}{4}\ln(1 - \frac{4}{3}p), \tag{2.2}$$

where p is the proportion of sites that are different.

2.1.4 Phylogenetic Methods

A phylogenetic tree is a binary tree in which leaves are associated with particular species or taxa, and the branching pattern of the tree reflects diverging evolutionary relationships. Branch lengths on the tree are associated with evolutionary divergence between sets of taxa. Molecular phylogenetics refers to a large collection of methods for inferring branching patterns from aligned molecular sequence data.[6] In general, the problem of finding an optimal tree associated with sequence data is NP-complete [60], however several approximate methods have been developed. The primary types of methods include maximum parsimony, distance-matrix methods, maximum likelihood (ML), and Bayesian inference. Maximum parsimony attempts to find the phylogenetic tree that minimizes the number of evolutionary changes required to explain the observed sequences. Distance-matrix methods first compute a matrix of pairwise distances between taxa and then find the tree that best approximates these distances. ML and Bayesian methods use specific models of evolution to assign probability distributions over trees. In this work we concentrate on distance-matrix methods because of their close connection with the finite metric spaces considered in applied topology.

2.1.4.1 Distance-Matrix Methods

Given a set of aligned molecular sequences, distance-matrix methods first compute the pairwise matrix of genetic distances using one of the metrics as described in Section 2.1.3.3. Then, the binary tree that best approximates those distances is iteratively fit to this data. This approach to phylogenetic inference were introduced by Cavalli-Sforza and Edwards in 1967 [28] and Fitch and Margoliash in 1967 [59]. The Fitch-Margoliash method uses a weighted least squares approach to tree-fitting, such that larger distances are weighted less, due to higher chances for random error. Distance-matrix methods are popular for their high speed and scalability as well as high accuracy in most cases.

[6] See Felsenstein's *Inferring Phylogenies* for a readable and thorough introduction to the field [58].

Data: $n \times n$ distance matrix D
Result: Phylogenetic tree on n leaves
while Tree not fully resolved ($n > 3$) **do**
  Compute Q matrix:
    $Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$;
  Identify pair of taxa i, j that minimizes $Q(i,j)$;
  Create new interior node u that joins i and j with edge length:
    $D(i,u) = \frac{1}{2}D(i,j) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} D(i,k) - \sum_{k=1}^{n} D(j,k)\right]$;
    $D(j,u) = D(i,j) - D(i,u)$;
  Create new $(n-1) \times (n-1)$ distance matrix where:
    $D(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)]$;
end

Algorithm 1: The Neighbor Joining Algorithm. Adapted from [142]

Currently, the most widely implemented distance-matrix method is neighbor-joining.[7] One particular reason neighbor-joining is popular is that under certain conditions, discussed below, it has been shown to exactly recover the correct tree. The neighbor-joining algorithm is a greedy approach to tree construction that iteratively joins the two closest nodes until a tree is fully resolved. The neighbor-joining algorithm is described in Algorithm 1.

2.1.4.2 Additive Metrics and the Four Point Condition

Arbitrary distance matrices are unlikely to admit a tree representation. Those that do are called *additive metrics*, because they can be represented as an additive tree. Additivity is the property that the distance between any two nodes will be equal to the sum of the branch lengths between them. A distance matrix admits a tree representation if and only if it is additive.

There is a straight-forward condition that must be satisfied for additivity, known as the *four point condition*. For a distance matrix to admit a tree representation,

$$d_{ij} + d_{kl} \leq \max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\} \tag{2.3}$$

for any four nodes $\{i, j, k, l\}$. The condition implies that there is a labeling on the four nodes

[7] Neighbor joining was introduced by Saitou and Nei in 1987 [121].

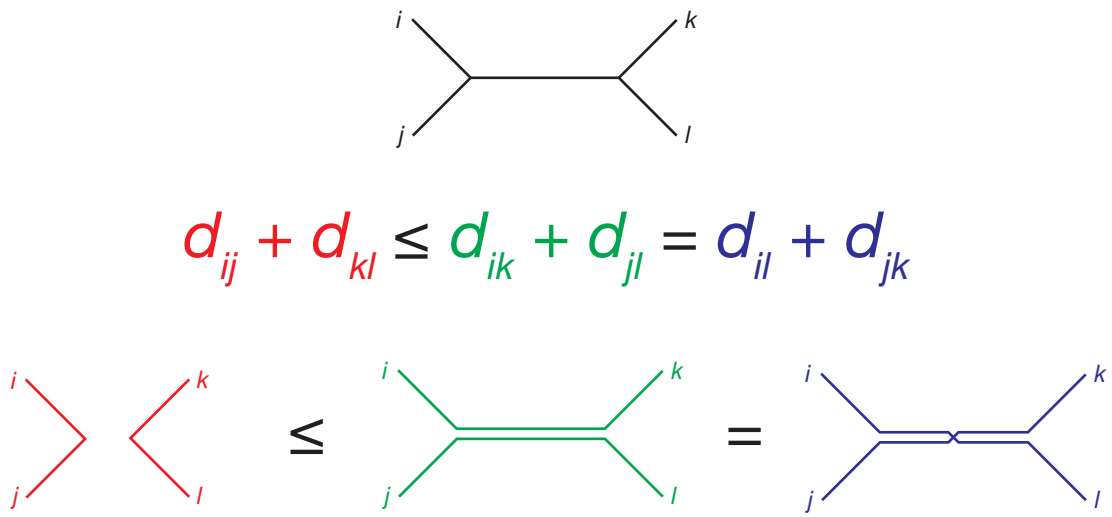


Figure 2.4: A visual interpretation of the four point condition for additivity. For any four leaves, there exists a labeling $\{i, j, k, l\}$ such that $d_{ij} + d_{kl} \leq d_{ik} + d_{il} = d_{il} + d_{jk}$. Of the three possible ways of arranging the sums of distances, two will involve traversing the internal branch, while one will involve only external branches.

such that

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}. \tag{2.4}$$

A visual interpretation of this condition is shown in Figure 2.4.

Sequence data can fail to be additive for several reasons. First, sequencing error. Errors can introduce noise into the measured genetic distances. Second, homoplasy. A homoplasy occurs when the same mutation is introduced multiple times in a set of organisms. The presence of homoplasy will underestimate genetic distance between taxa. Third, reticulate evolution. As described previously, in cases of reticulate evolution no tree will accurately describe the observed data. In this case, one can either attempt to find the tree that best fits the data, or search for an alternative representation of phylogenetic relationships.

2.1.4.3 Number of Tree Topologies

The number of unrooted bifurcating tree topologies with L leaves is $(2L-5)!!$.[8] This can be easily shown using induction. For $L = 3$, we have $\mathcal{T}(3) = 1$ and 3 branches. To pass to

[8]The double factorial is defined as $n!! = n(n-2)(n-4)\cdots$.

$L = 4$, we can add the fourth leaf to any of the 3 branches, resulting in 3 different topologies. For $L = 4$, we have $\mathcal{T}(4) = 3$. Every time we add a leaf, we add two branches – one external and one internal. For $L = n$, we have $\mathcal{T}(n) = (2n-5)!!$ and $2n-3$ branches. For $L = n+1$, we can add the new external branch to any of the current $2n-3$ branches. A rooted tree with L leaves can be considered as an unrooted tree with $L+1$ leaves. Therefore, the number of rooted bifurcating tree topologies with L leaves is $(2L-3)!!$ As can be seen, the number of tree topologies explodes with the number of leaves.[9]

2.1.4.4 The Space of Phylogenetic Trees

An unrooted phylogenetic tree with L leaves is characterized by its topology and the lengths of each branch. As shown in the previous section, there are $(2L-5)!!$ possible unrooted topologies. There are $2L-3$ total branches, of which L are external branches and $L-3$ are internal branches. Tree spaces refers to an abstract construction for representing each possible tree as a point in a geometric space. These studies were initiated by Andreas Dress and colleagues, who introduced a formalism known as *T-theory* (see [49, 48, 46]). We give here a brief flavor of these ideas; additional exposition can be found in [113, §7].

Consider a set of L taxa. A dissimilarity map is defined on L as $\delta : L \times L \to \mathbb{R}$, where $\delta(l,l) = 0$ and $\delta(l,m) = \delta(m,l)$. There are $\binom{L}{2}$ distances; the set of dissimilarity maps forms a vector space of dimension $\binom{L}{2}$. Furthermore, the set of all metrics will be the subspace of $\mathbb{R}^{\binom{L}{2}}$ that satisfies the triangle inequality. The space of trees is defined as the set $\mathcal{T}_L$ of dissimilarity maps that satisfy the four-point condition. The space can be logically decomposed into subspaces corresponding to a particular choice of topology. This will be taken as the union of $(2L-5)!!$ subspaces, each of dimension $2L-3$. Each subspace will have the structure of a metric cone in the space $\mathbb{R}^{\binom{L}{2}}$.

The geometric structure of this space was carefully studied by Billera, Holmes, and Vogtmann (BHV) in [15]. In that paper, the authors specifically considered rooted trees with

[9] As was observed by Walter Fitch, for 22 species there are on the order of Avogadro's number of topologies. ($N_{22} = 3.20e23$, $N_A = 6.02 \times 10^{23}$)

zero-length external branches, a space denoted as BHV_L, but the basic intuition generalizes to other types of trees. They defined a geodesic distance between trees of different topology and used it to define various metric properties on tree space. This analysis was extended by Zairis *et al.* in [149], in which unrooted trees with non-zero external branches were considered. The external branches are constrained to sit in the positive open orthant $(\mathbb{R}^{\geq 0})^L$. An evolutionary moduli space is then defined as the product

$$\Sigma_L = \mathrm{BHV}_{L-1} \times (\mathbb{R}_{\geq 0})^L. \tag{2.5}$$

The tree space construction allows one to define statistics, such as means and variances, on collections of trees in a meaningful way.

We show an example of the tree space construction on $L = 4$ and $L = 5$ leaves in Figure 2.5. The case of $L = 4$ is particularly simple to analyze. The metric cone is a subspace of $\mathbb{R}^{\binom{4}{2}=6}$. There are $(2*4-5)!! = 3$ tree topologies, corresponding to the patterns $((a,b),(c,d))$, $((a,c),(b,d))$, and $((a,d),(b,c))$. There are $(2*4-3) = 5$ branches: each topology will be a subspace in $\mathbb{R}^5$. The intersection of the subspace of each topology is a space in $\mathbb{R}^4$. The case of $L = 5$ also has a relatively simple structure. There are fifteen possible topologies, each with two internal branches. Each topology forms a hyperplane of dimension $\mathbb{R}^7$ Combinatorially, the topologies can be arranged as a Petersen graph. Intersections of three hyperplanes will correspond to degenerate cases with one internal branch is not resolved, as shown in Figure 2.5B. These facets sit in $\mathbb{R}^6$. It is important to think of the entire Petersen structure as being a cone, the origin of which is the 5-dimensional subspace consisting of only external leaves (see [15, Figure 14]).

Naturally, most data will not sit in $\mathcal{T}$. Whether or not this is simply due to noise or reflects reticulate processes will depend on the particular dataset. We can view the goal of phylogenetics as finding the best tree projection $\delta_T \in \mathcal{T}$ for arbitrary metric data X.

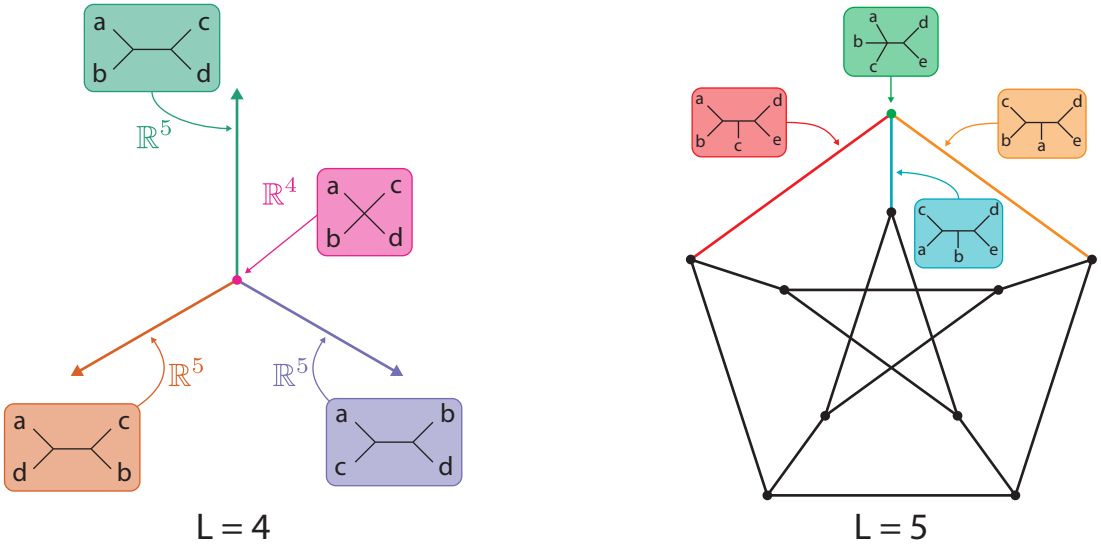


Figure 2.5: Examples of geometric representations of the space of trees on $L = 4$ and $L = 5$ leaves. (A) On four leaves the metric cone is a subspace of $\mathbb{R}^6$. There are three tree topologies, each of which corresponds to a 5-dimensional cone inside. The three topologies share a $\mathbb{R}^1$ facet corresponding to the degenerate topology.(B) On five leaves the metric cone is a subspace of $\mathbb{R}$. There are fifteen tree topologies, each of which corresponds to a 7-dimensional cone. The geometric structure of the space will map to a *Petersen graph*, as shown. There are 10 degenerate cases in which one internal branch is not resolved; these correspond to 6-dimensional facets, each joining three distinct topologies. The $n = 5$ subfigure is an adaptation of Figure 3.5 in [113, Ch 3].

2.1.4.5 Phylogenetic Networks

There are several existing methods for representing reticulate evolution. Most of these methods generalize phylogenetic trees into *phylogenetic networks*, which attempt to reconcile the presence of horizontal evolution in sequence data. However, most simply present corrections to phylogenetic trees, which can fail in cases where horizontal evolution is pervasive, as in many prokaryote datasets. Additionally, the resulting networks can be complex and difficult to interpret quantitatively. This can make it difficult to distinguish between phylogenetic incompatibilities due to noisy sampling and due to true reticulations. An example of a phylogenetic network using the split network approach is shown in Figure 2.6. Other methods include neighbor-net and median networks. Techniques such as phylogenetic networks and ancestral recombination graphs have been developed to describe reticulate evolution,

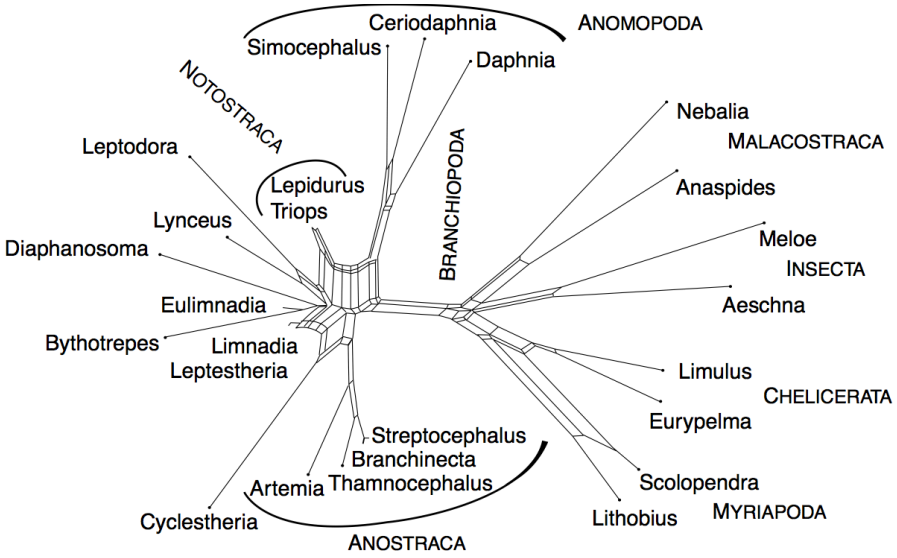


Figure 2.6: Example of a split network of genus Branchiopoda and outgroups. Computed using the Neighbor-Net algorithm. Phylogenetic incompatibilities are represented by conflicting splits. reprinted from BMC Evolutionary Biology 7:147 (2007).

but they have had only limited success due to difficulties of biological interpretation and computational infeasibility in all but the smallest datasets.

2.2 Topological Data Analysis

Topology is the branch of mathematics that formalizes our intuitive notions of shape. More concretely, topology provides the methods to characterize the properties of objects and spaces that remain invariant under continuous deformation. For example, squeezing a circle into an ellipse by compressing along one axis does not change the fact that the object encloses a single loop. Or, as we saw in the introduction, the coffee mug can be continuously deformed into the donut. Likewise, if we take a tree and change the lengths of its branches, the tree remains a tree.[10] In each of these examples, while the deformation has substantially altered *local* properties of the space, on a *global* level certain essential characteristics have remained unchanged. From the perspective of topology, the spaces are considered identical. The question then is how to formalize our idea of global shape in order to systematically reason about it.

Algebraic topology solves this problem by associating algebraic objects (an integer, for instance) that do not change under continuous deformation. These objects may capture properties like the number of connected components, the number of loops, and the number of holes in an object, and represent *topological invariants* of a space. Two spaces can only be deformed into one other if they share the same invariants. The circle and ellipse are identified as equivalent by the presence of a single loop. Neither can be deformed into a tree without introducing a cut, which would be a type of discontinuous deformation. Using these invariants, powerful ideas from abstract algebra can then be used to manipulate and reason about shape.

While topology has traditionally developed through the study of abstract spaces, leading to very rich and beautiful constructions[11], data does not come in the form of perfect continuous spaces. Recent effort over the past 15 years has focused on developing methods to apply

[10] It is important to draw a distinction between the notion of tree topology, in which the branch patterns determines the topology, and global topology, in which all trees are equivalent. While the former is more common in the phylogenetics community, here we consider the latter.

[11] For example, see the work of Thurston on low-dimensional topology

topology to real world problems in science and engineering. This work, collectively falling under the heading of *topological data analysis* (TDA), has focused on efficient algorithms for computing topological invariants from finite, noisy data. TDA now encompasses a wide range of efforts and can now be considered a branch of applied mathematics in its own right. It has emerged from substantial interdisciplinary effort between mathematicians, computer scientists, and domain experts.

In practice, a typical workflow for applying TDA to data is as follows. Data comes in the form of a set of n observations with p attributes, where p is often very large. The data is assumed to be a finite sample from some more complex space, from which we wish to infer either some sort of global structure or underlying model. The data is represented as a finite point cloud: a set of n points in p dimensions with an associated notion of distance. The point cloud is then transformed into a topological space by associating different sets of points with each other. The associations can be constructed in different ways – for instance, one of the most common constructions associates points within a certain distance d from one another. Computational approaches are then used to measure informative topological properties from the space.

In this thesis, we use methods from TDA to study problems in evolutionary biology and genomics. Our data is typically aligned genomic sequences from sets of related organisms. If our sequences are each of length L, then we can imagine our data as points in an L-dimensional sequence space. A genetic sequence metric, such as the Hamming metric, measures distance.

The two main methods from TDA that we employ are *persistent homology* and *mapper*. Persistent homology provides a way to efficiently compute the topological invariants of a space across multiple scales, while mapper provides an approach for condensed representation and visualization of high-dimensional data. In this section, we provide an overview and discussion of these two methods from the perspective of an end-user, treating each method as a pipeline for transforming from raw data to a concise topological summary. While the

mathematical literature on these methods is extremely deep, our goal is to explain things in sufficient detail for a wide audience to grasp the main ideas. We therefore include a brief introduction of the basic mathematical concepts we employ. The primary concept we require is *homology*, a particular way in which topological invariants can be assigned to spaces.

The following sections draw on several excellent reviews of TDA, including [25], [51], and [63]. A more thorough introduction to algebraic topology can be found in [69].

2.2.1 Preliminaries

As stated above, our data is a set of n points, $S = \{s_1, \ldots, s_n\}$. Each point is a vector with p features, $s_i = (s_{i1}, \ldots, s_{ip})$. We refer to the collection of points, embedded in a space with an appropriate metric structure, as a point cloud. We wish to associate a collection of algebraic objects to the point cloud in order to quantify its shape. To do so, our first step is to construct a topological structure on top of the point cloud, called a *simplicial complex*. The structure will consist of a set of simplices pieced together in such a way that they approximate the shape of the point cloud in a sensible way. Shape is then quantified using the notion of *homology*. This section provides the definitions necessary to understand homology.

2.2.1.1 Simplices and Simplicial Complexes

The building blocks of our topological structures are simplices. A *simplex* is something like a point, a line, a triangle, or any higher-dimensional generalization of such. Formally, a k-simplex is a k-dimensional polytope which is the convex hull of $k+1$ vertices, as shown in Figure 2.7. A simplex can be represented by its list of vertices, i.e. $\sigma = (s_1, s_2, s_3)$. An m-face of a simplex is the space spanned by the set of $m+1$ vertices, and is itself a simplex. For example, the 0-faces and 1-faces of a simplex are its vertices and edges, respectively. The $(k-1)$-faces (faces of co-dimension 1) of a k-simplex are called facets. Facets are represented as $\sigma_{(-i)}$, which implies the facet generated by elimination of the i-th vertex.

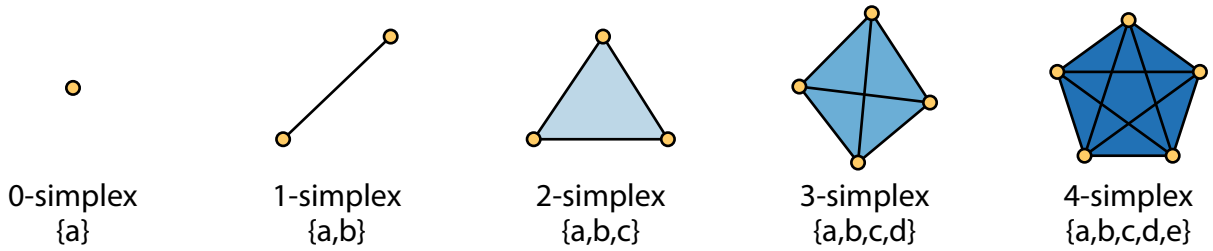


Figure 2.7: Simplices are the fundamental building blocks of our topological structures. They can be thought of as triangles generalized to arbitrary dimension. Here we show k-simplices for $k = 0$ to $k = 4$.

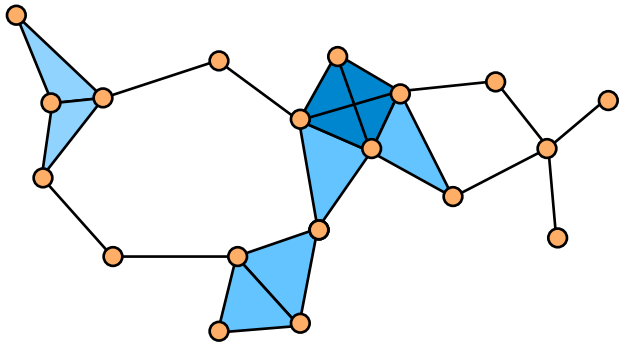


Figure 2.8: A finite simplicial complex is an object built from finite number of simplices glued together in a sensibly nice way.

A *finite simplicial complex* K is built on the vertex set S from simplices glued together in such a way that (1) any face of a simplex in K is also in K, and (2) the intersection of any two simplices in K is a face of both simplices. An example of a simplicial complex is shown in Figure 2.8.

We are interested in combinatorial operations that can be performed on a simplicial complex K. In general, these operations will act on subsets of simplices of fixed dimension k. These subsets are called *k-chains*, and can be represented as formal sums $C_k = \sum_j \alpha_j \sigma_j$. The coefficients α_j will be taken to be over $\mathbb{Z}_2$ (i.e. 0 and 1). Two consequences of this choice are (1) $\sigma + \sigma = 0$, and (2) we consider simplices without regard to orientation.[12]

[12]In general, an algebraic topology can be defined with coefficients in arbitrary fields. We use $\mathbb{Z}_2$ for simplicity, efficiency, and because properties, such as torsion, that arise over more complex fields are not expected to be present in the biological data we consider. It is important to keep this in mind, as it was in fact shown that torsion can arise in real data in [27]. In that paper, an association was shown between the

An important operator is the boundary operator, $\partial : C_k \to C_{k-1}$. The boundary of a simplex σ, $\partial_k \sigma$, is the sum of its facets.

$$\partial_k \sigma = \sum_i \sigma_{(-i)} \tag{2.6}$$

The boundary of a chain is $\partial C = \sum_j \partial \sigma_j$. As a simple example, consider the 2-simplex $\triangle$ defined by vertices $\triangle = (a, b, c)$. We have $\partial \triangle = (a, b) + (b, c) + (a, c)$. Further, we have $\partial \partial \triangle = 2(a) + 2(b) + 2(c) = 0$. In fact, the property $\partial \partial C = 0$ will hold for any chain C.

We can additionally define more refined types of chains. A *cycle* is a chain with empty boundary, $\partial C = 0$. A *boundary cycle* is a k-cycle that is the boundary of a chain in dimension $k + 1$.

We can use these definitions to construct various groups on a simplicial complex K. The set of all chains of dimension k forms the chain group C_k. The set of all cycles of dimension k forms the cycle group Z_k. The set of all boundary cycles of dimension k forms the a group B_k. The latter two groups can be understood in terms of the boundary operator ∂ acting on K. The group Z_k is the kernel of the boundary operator, $Z_k = \ker \partial_k$. That is, it is the set of all k-chains that are sent to 0 by the boundary operator. The group B_k is the image of the boundary operator, $B_k = \operatorname{im} \partial_{k+1}$. That is, it is the set of all k-chains which are themselves the boundary of $(k + 1)$-chains in K. These groups have a particularly simple relationship to one another which is shown in Figure 2.9.

2.2.1.2 Homology

We are now ready to define homology, which will allow us to discuss and compare shape in a quantitative way. The j-th homology of a simplicial complex K is defined as the quotient group

$$H_j(K) = Z_j / B_j = \ker \partial_j / \operatorname{im} \partial_{j+1}. \tag{2.7}$$

---

space of natural images and the Klein bottle.

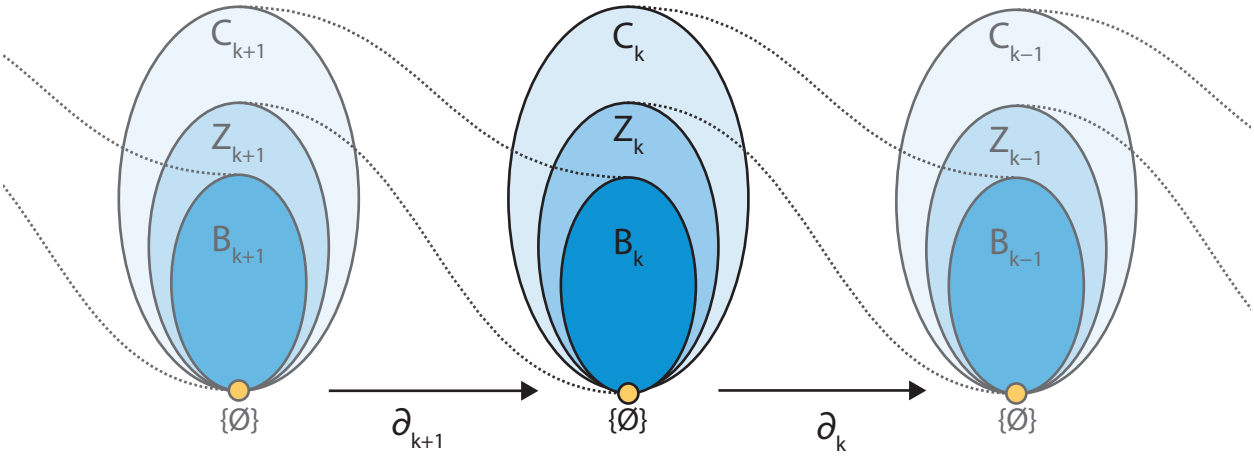


Figure 2.9: Relationship between the chain group (C_k), cycle group (Z_k), and boundary group (B_k). Specifically, $B_k \subset Z_k \subset C_k$. We show the action of the boundary map ∂_k on each group at each dimension. Of particular note are the relations $\partial_k : C_k \to B_{k-1}$ and $\partial_k : Z_k \to \emptyset$. Figure adapted from [56].

In words, homology is the group generated by equivalence classes of the cycle group Z_j, where equivalence is defined up to B_j. Elements of the homology group are classes of homologous cycles. Two j-cycles are homologous is they differ by the boundary of a $(j+1)$-chain. We work through a simple example in Figure 2.10.

The rank of the homology group $\|H_j(K)\|$ is the Betti number b_j. Intuitively, the Betti number represents the number of j-dimensional holes in the simplicial complex.

2.2.1.3 Constructing Complexes From Data

Finally, we must consider how to construct a simplicial complex from a given point cloud S.[13] There are two common constructions we will describe: the *Čech complex* and the *Vietoris-Rips complex*. Both constructions involve a scale parameter ϵ, and balls of radius ϵ placed at the center of each vertex in S. Edges are drawn between vertices when balls overlap, that is, when $d(v_a, v_b) < 2\epsilon$. Where the two constructions differ is in how higher-dimensional simplices are filled in.

[13] In fact, this step is arguably the most important step in applying a TDA pipeline.

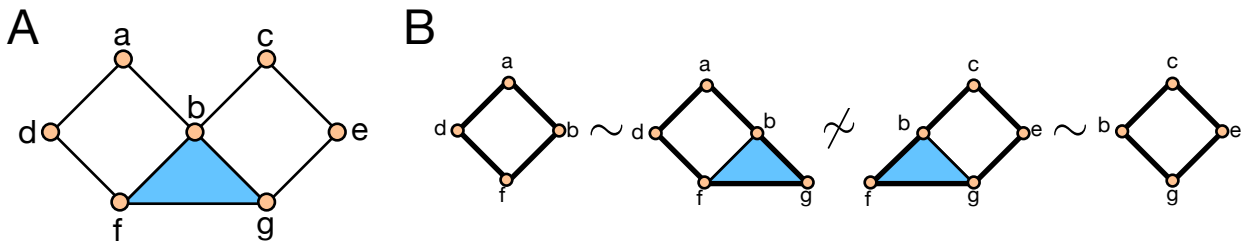


Figure 2.10: (A) A simplicial complex defined on a set of 7 vertices, $S = \{a, \ldots, g\}$. The object has one connected component ($b_0 = 1$) and two holes ($b_1 = 2$). (B) Four cycles that can be defined on the complex. Cycles $z_1 = \{(a,b) + (b,f) + (f,d) + (d,a)\}$ and $z_2 = \{(a,b) + (b,g) + (g,f) + (d,f) + (d,a)\}$ are homologous, differing only by the cycle $c_1 = \{(b,g) + (g,f) + (f,b)\}$ which is the itself the boundary of the closed triangle (b,f,g). Likewise with cycles z_3 and z_4. The two sets of cycles are not homologous with each other, and there are therefore constitute two independent elements of the homology group $H_1(S)$. Note that the basis is not unique.

The Čech complex consists of the set of simplices σ with vertices $s_1, \ldots, s_k \in S$ such that

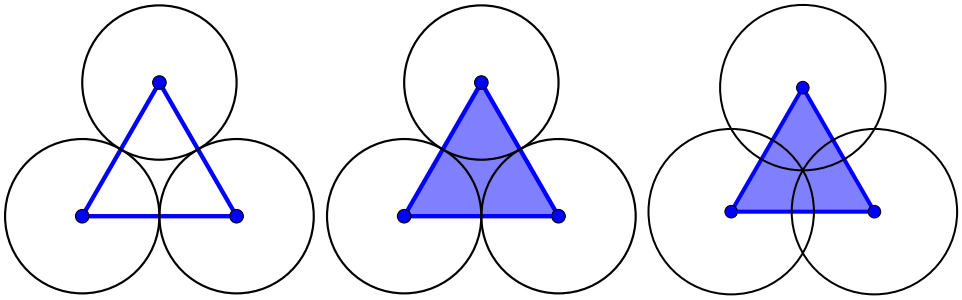
$$\text{Čech}(S, \epsilon) = \{\sigma \in S \mid \cap_i B(s_i, \epsilon) \neq 0\}. \tag{2.8}$$

That is, the simplex $\sigma_{(s_x,\ldots,s_z)}$ is present if the intersection of balls of radius ϵ centered on vertices $(s_x, \ldots, s_z)$ is nonempty. The Vietoris-Rips complex, $VR(S, \epsilon)$, is defined as

$$\text{VR}(S, \epsilon) = \{\sigma \in S \mid \text{diam}(\sigma) \leq 2\epsilon\} \tag{2.9}$$

where $\text{diam}(\sigma) = \{\sup d(i,j) \mid i, j \in \sigma\}$. In the Vietoris-Rips complex, a higher-dimensional simplex is filled in if every pairwise distance is less than 2ϵ. The difference between the two constructions is shown in Figure 2.11. In general, $\text{Čech}(S, \epsilon) \in VR(S, \epsilon)$.

The Čech complex is theoretically preferable because it comes with a *Nerve theorem*, which essentially states that the topology of the resulting complex will be equivalent to the topology of the union of balls used to create it. However, the Čech complex has drawbacks that prevent it from being widely applied to arbitrary data. First, computing the intersection of arbitrary balls is an expensive operation. While efficient algorithms exist in Euclidean space (the miniball algorithm [62]), it is much more difficult in arbitrary metric spaces. Furthermore, the Čech construction explicitly requires an ambient space in which the data



(a) Three Points (b) Vietoris-Rips Complex (c) Čech Complex

Figure 2.11: An example of the difference between Vietoris-Rips and Čech complex on an equilateral triangle. Consider each point to be 1 unit apart. In the Vietoris-Rips complex, the triangle is filled in when every pairwise edge is connected ($\epsilon = 0.5$). In the Čech complex, the triangle is only filled in when all three balls intersect ($\epsilon = 0.577$).

is embedded. For data which comes in the form of a finite metric space, it may not be clear what is the ambient space.[14] In practice, the Vietoris-Rips complex is more widely applied, because it requires only the set of pairwise distances between each vertex and a scale parameter ϵ. The complex can be directly read off from the set of edges (known as the *1-skeleton*), making it extremely fast to compute.

2.2.2 Persistent Homology

Persistent homology is a tool developed under the umbrella of TDA that allows us to computationally study the shape of a point cloud across multiple scales simultaneously. Shape is quantified in terms of topological invariants representing homology, as discussed in the previous section. To understand why multiscale information might be of interest, consider the example in Figure 2.12. The data is sparse and noisy, but, to the eye, immediately appears to consist of two circles joined at a point along their edges. The two circles, however, are of a different radius. In the Figure, we show Vietoris-Rips complexes constructed at

[14]This is indeed the case for the genomic data we consider: it is not immediately obvious what the intersection of three sequences defined over a finite alphabet should be. We discuss this further in Chapter XX.

different scale parameters on the data. We observe that while some scale parameters are sufficient to resolve one or the other of the two circles, no single scale parameter is sufficient to simultaneously capture the two shapes. Persistent homology solves this by providing a way to track the shape information across *all* scale parameters.

Our object of study is the nested set of simplicial complexes, called a *filtration*, that is produced by tuning the scale parameter up to some threshold. At the smallest scale, $\epsilon = 0$, the complex consists only of disconnected points. As the scale parameter is increased, the topology of the complex changes – clusters merge, holes and loops form, other holes and loops are filled – until the complex is fully connected. Each aspect of shape represents a topological invariant, and as the scale is changed, the birth and death of different invariants encoded as an interval (b_i, d_i).

The shape information can be concisely summarized in a *barcode diagram*. The barcode diagram represents topological features as horizontal line segments, annotated with a birth-death interval, and a dimension. The birth time is when a particular invariant first appears in the complex, and the death time is when the invariant is collapsed in the complex. Shape information by dimension. H_0 represents the number of connected components and is roughly equivalent to a hierarchical clustering of the data. Higher dimensions represent loops (H_1), voids (H_2), and their generalizations in the data. The number of bars at a particular scale will be the Betti number b_n for that complex. Taken together, the barcode diagram represents a complete and quantitative picture of the shape of the data.

The information can be equivalently represented as a persistence diagram, which is a scatter plot of invariants with birth time on the x axis and death time on the y axis. The barcode diagram and persistence diagram for the two circles data is shown in Figure 2.13. First, looking at H_0, we see that the data begins disconnected and becomes connected at around $\epsilon = 24$. Next, looking at H_1, we count eight loops across a range from $\tilde{5}$ to $\tilde{8}0$. Two of these loops persist for what appears to be an appreciable length of time. We associate these two loops with the two circles that we identified qualitatively from the raw point cloud

data.

The intuition behind persistent homology is exactly that: somehow the good or interesting features will be robust and persist over long scales. In the barcode diagram, this corresponds to longer bars; in the persistence diagram, this corresponds to points sitting far from the diagonal. Invariants that we observe persisting for only short scales are likely to be noise or other artifacts.[15] And because a single scale is not capable of representing all features of the data, we examine all scales simultaneously.

In fact, the persistence algorithm is more powerful than that, and can return not only the intervals associated with the invariants, but *representative cycles* of each invariant. The representative cycles correspond to the set of simplices that surround an invariant, and can be used to determine which data points are somehow involved in a particular invariant.

To summarize, a complete description of the persistent homology pipeline is shown in Figure 2.14. The pipeline is as follows: A dataset, $S = (s_1, \ldots, s_N)$, is represented as a point cloud in a high-dimensional space (not necessarily Euclidean). From the point cloud, a nested series of simplicial complexes, or a filtration, is constructed, parameterized by a filtration value ϵ. The filtration is represented as a list of simplices defined on the vertices of S, annotated with the ϵ at which the simplex appears. Given a filtration, the persistence algorithm is used to compute homology groups. The 0-dimensional homology (H_0) represents a hierarchical clustering of the data. Higher dimensional homology groups represent loops, holes, and higher dimensional voids in the data. Each feature is annotated with an interval, representing the ϵ at which the feature appears and the ϵ at which the feature contracts in the filtration. These filtration values are the *birth* and *death* times, respectively.

There is another way of applying persistent homology. This is the level-set filtration, which we now describe. To do.

As primarily end-users of persistent homology, the details of the persistence algorithm

[15] The obvious question of how to rigorously determine what makes a good interval is an open question that is currently being addressed by a number of different groups.We discuss this further in Section 2.2.2.2.

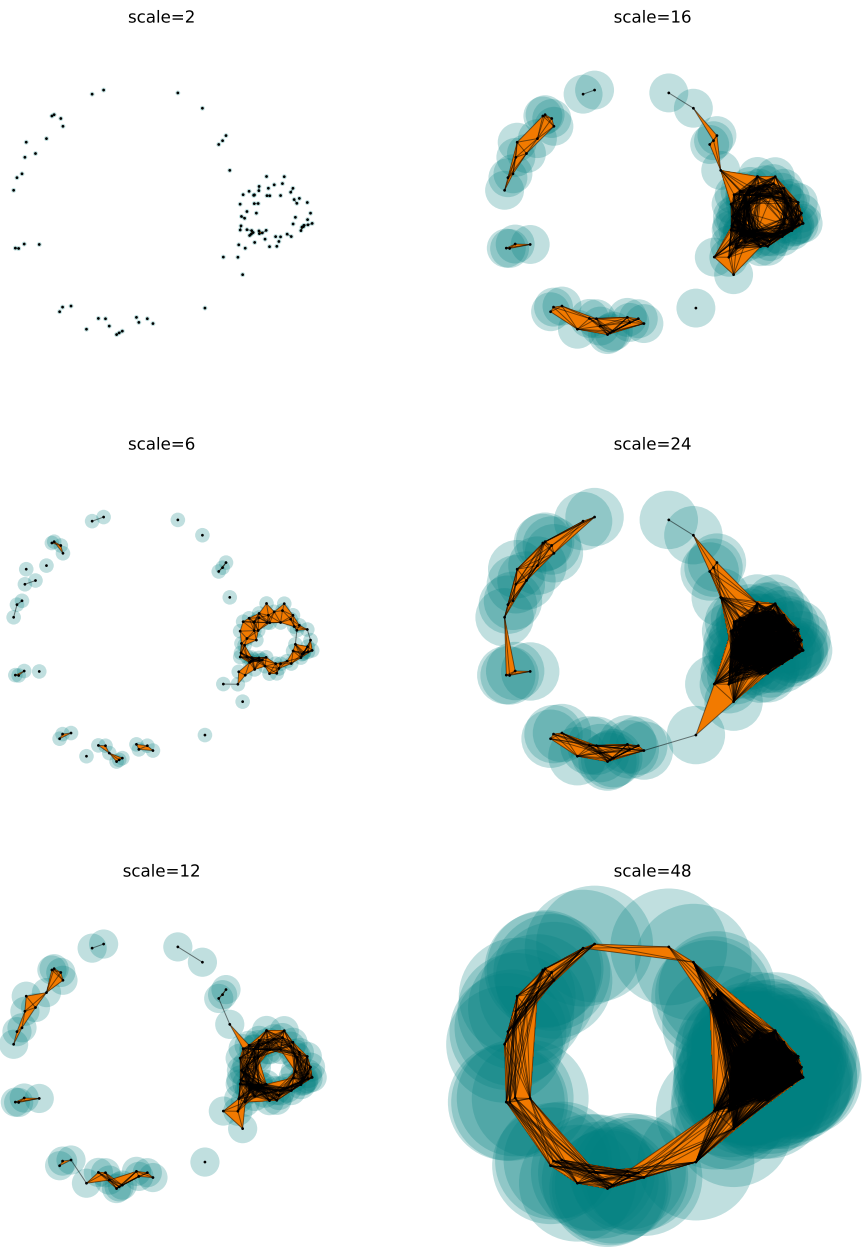


Figure 2.12: An example of constructing a filtration. The nested series of complexes form a filtration. Persistent homology will compute and track the homology at each scale. Adapted from Lesnick.

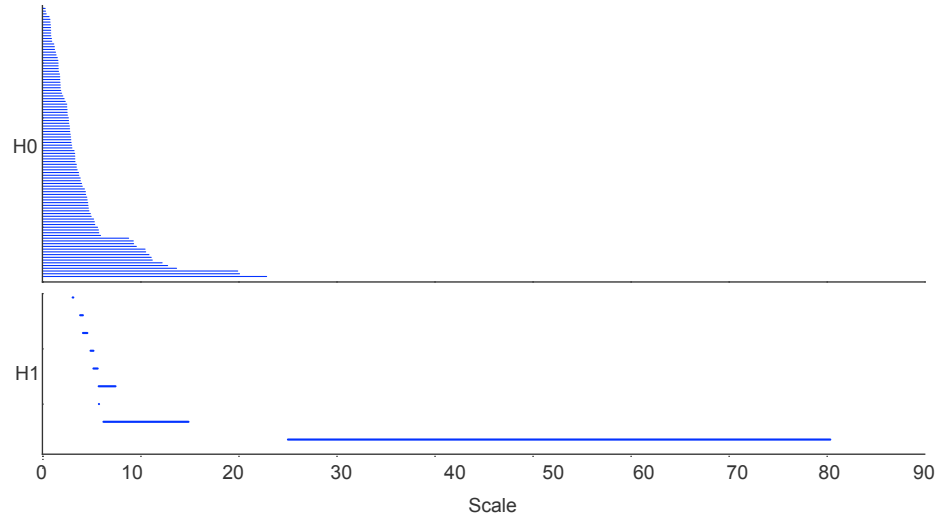


Figure 2.13: Barcode Diagram for the example in Figure 2.12. H_0 represents connectivity. H_1 represents the holes. Two holes are present at two different scales. Also some noise in the data.

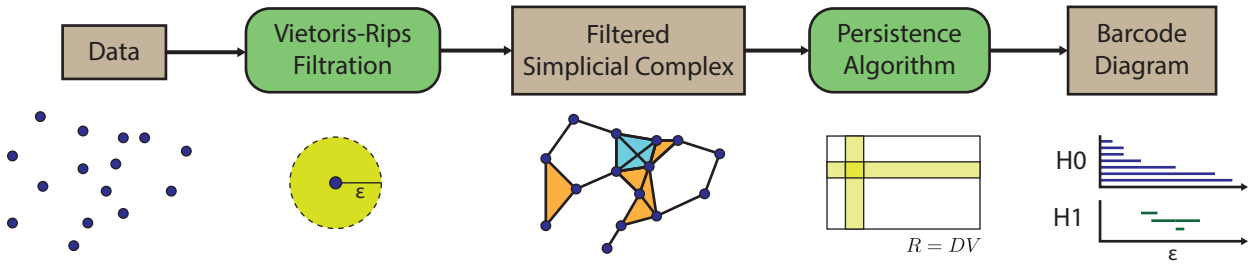


Figure 2.14: The Persistence Pipeline.

are largely beyond the scope of this thesis. Effectively, it involves manipulating the boundary matrix into a particular reduced form, from which each bar and representative cycle can be read off. Several packages for computing persistent homology have been developed, including Javaplex [130], Dionysus [103], Perseus [105], Gudhi [100], and PHAT [14]. Additionally, there is a TDA package for R which wraps functions from Dionysus and Gudhi in a user-friendly frontend [57].[16]

[16]In our work we have relied on a variety of these packages. For straight-forward construction of the barcode diagram, we find the R package TDA easiest to use. If one needs to directly build and manipulate filtered simplicial complexes, Dionysus has convenient Python bindings. For large datasets, PHAT and it's parallel implementation DIPHA [12, 13] are recommended.

2.2.2.1 Stability of the Persistence Algorithm

While not directly utilized in this thesis, the stability statement is important to include because it gives a solid grounding for comparing the persistence diagrams from different data. Of particular interest, it gives conditions on the effect of noise on data sampled from a particular object. The stability result guarantees that small perturbations in the input data will produce only small changes in the output diagrams. The result has many formulations, but is due originally to Chazal *et al.* in [30]. The statement requires two things: (1) a notion of distance between the input data D and the perturbed data D', and (2) a notion of distance between the resulting diagrams B and B'.

We need a notion of distance between persistence diagrams. Recall the persistence diagram consists of a set of intervals (b_i, d_i) along with the diagonal. First we will need the concept of a *matching*. For two persistence diagrams B and B', a matching is simpling a mapping of intervals in B to intervals in B', where we allow points to match to the diagonal (to account for cases with an unequal number of points). For a matched pair of intervals (a, b), we define the L_∞ distance as

$$d_\infty(a, b) = \max\{|a_x - b_x|, |a_y - b_y|\}. \tag{2.10}$$

The *bottleneck cost* of a matching between two diagrams is defined as the maximum L_∞ distance among all matched pairs. The *bottleneck distance* is defined to be the minimal bottleneck cost across all matchings,

$$d_B(B, B') = \inf_{\gamma \in \Gamma} \sup_{p \in B} \|p - \gamma(p)\|_\infty \tag{2.11}$$

The matching with minimal bottleneck cost is the *bottleneck matching*.

We need a notion of distance between finite metric spaces. Here we will use the *Gromov-Hausdorff distance*, which measures how far apart two spaces are from being isometric. It measures the longest distance from a point in one set to the closest point in another set within a metric space.

$$d_{GH}(X, Y) = \inf_{f,s} d_H(X, Y) \tag{2.12}$$

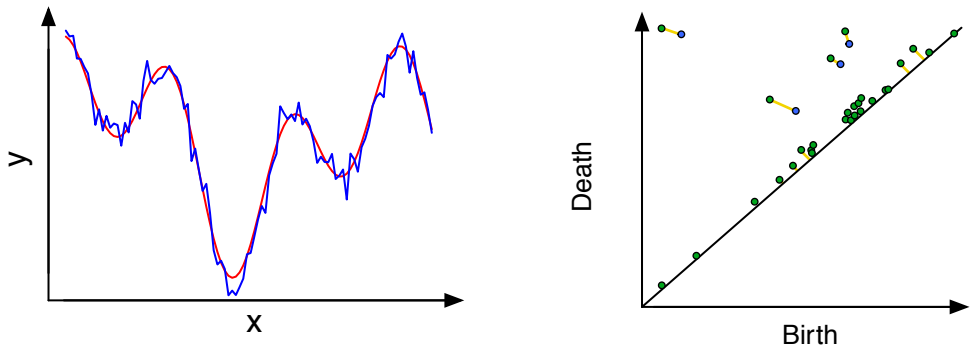


Figure 2.15: An example of the stability of the persistence diagram with respect to noise. (A) A function (red), and a noisy sampling of the same function (blue). (B) The level-set persistence diagrams for the two samples. A bottleneck matching is indicated with yellow lines. The stability result gives an upper bound on how much the diagrams can diverge as the sample diverges from the true function. This figure is adapted from [51].

The result of [30] states that the bottleneck distance between B and B' is bounded by the Gromov-Hausdorff distance between the finite metric spaces embedded in A and B.

$$d_B(H_K(X), H_K(Y)) \leq d_{GH}(X, Y) \tag{2.13}$$

The idea is easiest to visualize using a level-set persistence example, which we show in Figure 2.15. Here we see a simple function, $f(x)$, and a noisy sample of the same function. The persistent diagrams of each are shown in Figure 2.15B, along with a bottleneck matching, shown in yellow. As can be seen, most of the noise introduced is reflected in intervals close to the diagonal. While this example was for a simple level-set filtration, the result has an extension to the arbitrary finite metric spaces (X, d_X) which we primarily consider in this thesis.

2.2.2.2 Statistical Persistent Homology

Persistent homology has been developed largely as an exploratory tool. However, one would like to integrate it in data analysis pipelines more broadly, which requires some notions of statistics to be developed. One of the difficulties is that the persistence diagram, consisting of a multiset of points in the plane, can be somewhat unwieldy to work with. Substantial recent

work in the TDA community has focused on these questions in order to develop statistical foundations for persistent homology. We give here a brief flavor of some of these ideas and their relation to our own work. There are three main threads in statistical persistent homology:

1. Functional summaries of the persistence diagram

2. Confidence intervals on persistence diagrams

3. Probability measures on the space of persistence diagrams

First, confidence intervals on the persistence diagram. This addresses the question of when a bar is significant topological feature and when it can be considered topological noise. Fasy *et al.* have developed ways of generating confidence intervals for persistence diagrams [56]. They use a filtration on a kernel density estimation of the data, and bootstrap resampling, to put a line off the diagonal below which points are to be considered noise. Can be used to handle outliers and noise in the data. A few approaches: bootstrap estimates and also density bifiltrations. Using a bootstrap subsampling approach is able to tighten this confidence set. See the example of a circle with outliers in Figure 2.16. Only using the density estimate with a bootstrap estimator is capable of recognizing the circle as a significant feature. Subsampling estimates were studied further in [31]. Related approaches were developed by Blumberg *et al.* in [16].

Second, functional summaries of the persistence diagram. From functional summaries, machine learning approaches can be used downstream. Bubenik has developed the language of persistence landscapes [22, 21]. Essentially, the landscape is generated by rotating the persistence diagram 45 degrees and dropping a tent at each point. The silhouette is taken

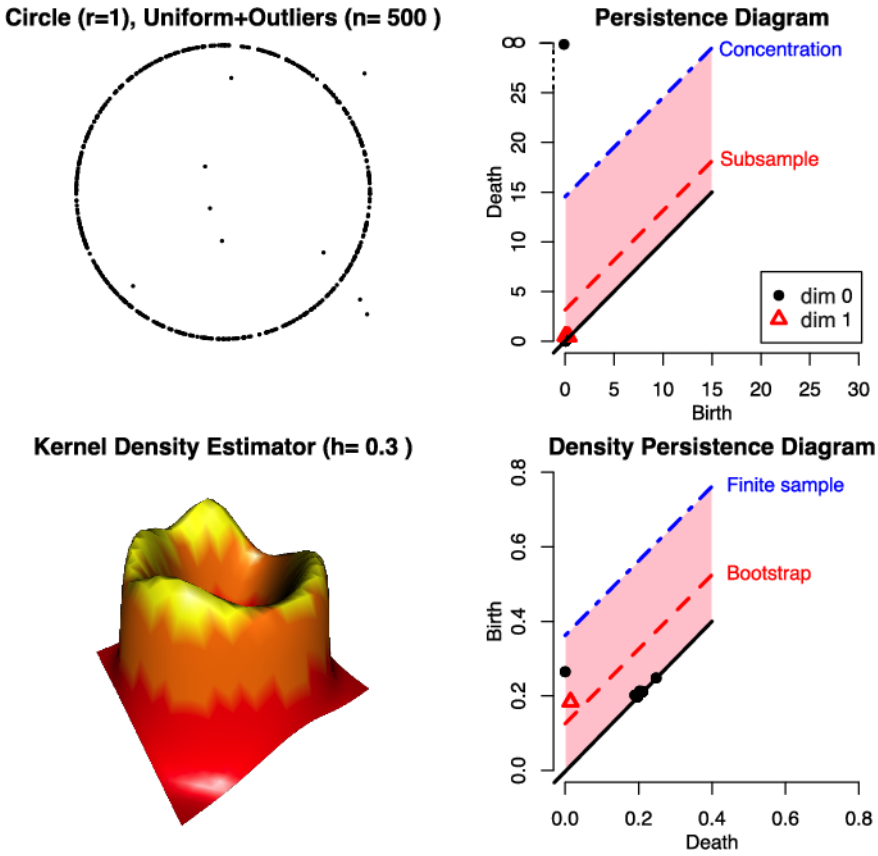


Figure 2.16: An example of how statistical persistent homology can be used to handle both noise in data and put confidence sets on the persistence diagram. Figure is taken from [56].

by weighting the contribution of each point based on its height.

$$\Lambda_p(t) = \begin{cases} t - x + y & t \in [x - y, x] \\ x + y - t & t \in (x, x + y] \\ 0 & \text{otherwise} \end{cases} = \begin{cases} t - b & t \in [b, \frac{b+d}{2}] \\ d - t & t \in (\frac{b+d}{2}, d] \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

An example of the persistent landscape and silhouette is shown in Figure 2.17.

A second approach has been taken by Schweinhart and MacPherson in [98], in which a transformation of the intervals is used to characterize random polymers in terms of a fractal dimension. A related approach has recently been proposed by Kwitt *et al.* that represents the persistence diagram in a kernel space for use in machine learning [86, 117]. We explore the use of functional summaries of the barcode diagram in a statistical inference setting in Chapter 4.

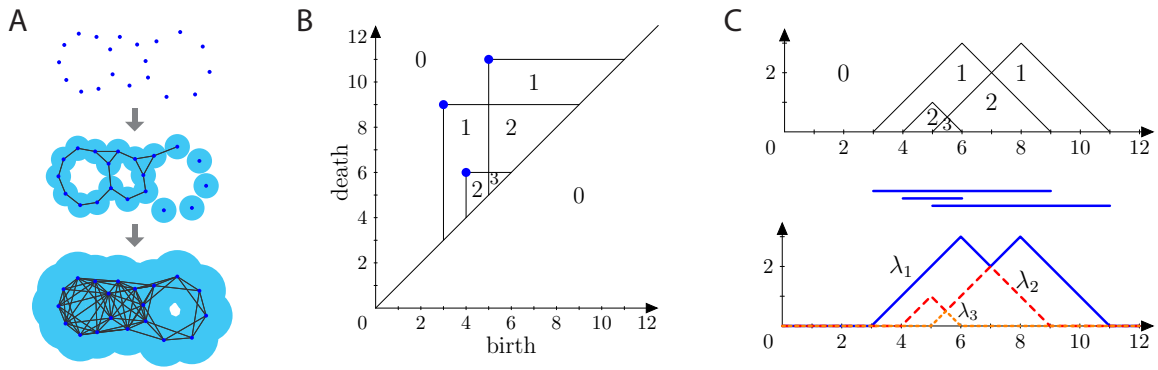


Figure 2.17: Persistent Landscapes. (A) A simple filtration on a set of points. (B) The persistence diagram from this data. (C) Transformed into a persistence landscape. This figure is adapted from [21, Fig. 2].

Finally, probability measures on the space of persistence diagrams. Several authors have examined the space of persistence diagrams as a Polish space, with notions of mean and variance. A Polish space has a well defined notion of mean and variance. See the work of Turner [134] and Mileyko [102]. This work crucially relies on distances between diagrams that are based on matchings, as discussed in Section 2.2.2.1. Further work has focused on statistical properties of the persistence diagrams themselves, such as [32] Establishing these foundations would lead to directly using the persistence diagram in statistical applications, but as of yet no efficient tools have been developed.

2.2.3 Mapper

Mapper is an approach for the representation and visualization of patterns in high-dimensional data. As such, it sits within the larger class of dimensionality reduction algorithms for exploratory data analysis (EDA), such as multidimensional scaling (MDS) [85], Isomap [131], and t-SNE [135]. A perspective on how these various approaches to EDA are related is shown in Figure 2.18. The primary distinction between Mapper and the existing class of nonlinear dimensionality reduction algorithms is that Mapper seeks to preserve the topology of the input data, rather than geometry. Compared to existing approaches, Mapper has the

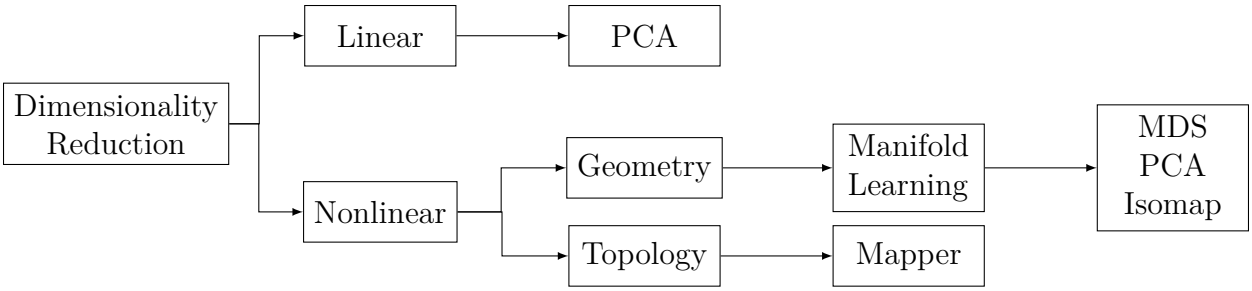


Figure 2.18: Dimensionality Reduction Approaches for Exploratory Data Analysis.

following advantages: (1) it is coordinate free, depending only on the metric properties of the data; (2) an invariance to deformation, which provides a robustness against noise; and (3) results in a compressed representation, which gives the ability to handle extremely large data. Mapper was initially developed by Singh and Carlsson in [123]; further exposition and examples can be found in [96]. Mapper has been applied to problems in RNA folding [19], breast cancer subtype classification [110], and genetic associations in type 2 diabetes [91].

A simple example of the Mapper algorithm is shown in Figure 2.19, which is adapted from [96]. As input, a point cloud X with an associated metric, Euclidean or not (Figure 2.19)A. First, a filter function is applied to the data (Figure 2.19B). The filter function maps the original points onto the real line, $X \to \mathbb{R}$. Standard filters include things like the mean, density, L_1-centrality, and the first and second components of a PCA decomposition. Second, the projected space is split into overlapping bins based on a resolution (bin size) and overlap parameter.[17] Third, the bins are then clustered in the original high-dimensional space (Figure 2.19). Each cluster will form a node in the graph representation.[18] Finally, nodes that share points in the original space are connected by an edge.

Our use of Mapper will be primarily as a means of visualizing relationships in sequence data. While the resulting graphs cannot be strictly interpreted in a phylogenetic sense, they will provide valuable information about evolutionary relationships. We use the commercial

[17]Multiple filter functions can be used and binned on a grid.

[18]Nodes in a Mapper graph consist of multiple points in the original data; this is essence of the compressed representation.

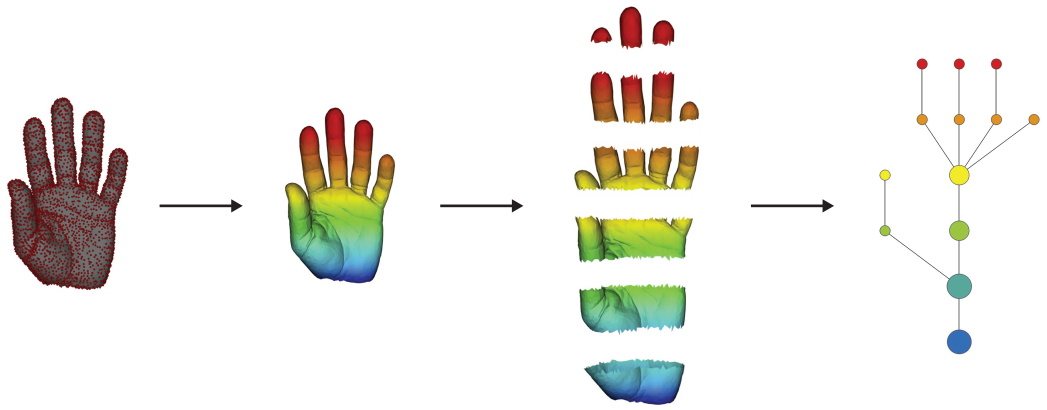


Figure 2.19: Mapper starts with a set of data points and a filter function f and produces a colored graph that captures the shape of the data. (A) The image of the function f is subdivided into overlapping intervals. (B) Each piece is clustered separately. (C) Each cluster is represented by a colored disk: a bin of points. The color of each bin corresponds to the average value of the filter function f on the data points inside the bin. (D) Identify pairs of bins that have points in common and (E) connect pairs of bins that have points in common by an edge.

implementation of Mapper developed by Ayasdi [7]. An open-source implementation of Mapper is available in the Python Mapper package [104].

2.3 Applying TDA to Molecular Sequence Data

Aligned molecular sequence data can be naturally viewed as a point cloud in a high-dimensional space, which we loosely call *sequence space*. The particular structure of sequence space will be determined by the length, L, of the aligned sequences, and the alphabet, Q, over which the sequences are defined. The typical sequence alphabet will be either nucleotides or amino acids. The dimension of the space is determined by L. Sequence space will therefore consist of the $||Q||^L$ possible sequences. Together with any of the standard genetic distance measures, this forms a metric space.

The processes of evolution can be seen as an exploration of sequence space. An individ-

ual genotype can be assigned a fitness which will describe the probability of reproductive success in a particular environment. Clonal evolution is the process of smoothly moving through sequence space, while reticulate evolution is the process of making discontinuous jumps through the space. Our data consists of a subset of points sampled from sequence space. These points reflect a particular evolutionary history. As more data is acquired, regions of sequence space will become more densely sampled and our ability to reconstruction evolutionary relationships will be improved.

Given sequence data, our program is to (1) encode the data as a finite metric space, (2) use tools from TDA to characterize the topology of the data, and (3) interpret the topology in an evolutionary context. In particular, we apply persistent homology, and read phylogenetic information contained in the dataset off the resulting barcode diagram. This idea was first proposed in [29]. In that paper, the authors developed two metrics for measuring reticulate evolution using homological features: (1) topological obstruction to phylogeny (TOP), which uses the L_∞-norm of the barcode as a coarse measure of reticulation; and (2) irreducible cycle rate (ICR), which uses temporal annotations to measure the average number of H_1 features per unit time. They applied this approach to a variety of viral datasets, including influenza and HIV. The work presented in this thesis extends this work in several substantial ways. We make two preliminary remarks before considering a more complex example.

2.3.1 Topology of Tree-like Metrics

An important foundational point was demonstrated by Carlsson in [29]. Recall that tree-like data will have an additive metric, as described in Section 2.1.4.2. In [29], it was proven that for additive metric spaces, the Vietoris-Rips filtration of the data will consist of a nested set of acyclic complexes. Consequently, the persistent homology of additive data will have nontrivial topology only in dimension zero. Furthermore, while noise in the data will introduce small deviations from additivity, the theorem puts bounds on the size of the topological features that can arise in this manner. These bounds rely on the Gromov-Hausdorff stability

conditions described in Section 2.2.2.1. On the other hand, if the evolutionary history includes reticulate events that cannot be represented as a tree, these events will be captured as non-trivial higher dimensional homology in the barcode diagram, an idea which we develop below. This theorem provides an important negative control in using TDA to characterize reticulate evolution.

2.3.2 The Fundamental Unit of Reticulation

In population genetics, there is a simple test for the presence of reticulate evolution in sequence data called the *four-gamete test* [73]. The test assumes only an infinite-sites model, which states that for a sufficiently long genome, a particular residue can only ever undergo a single mutation. Put another way, there is no multiple-mutation or back mutation. The infinite-sites model has three consequences: first, one need only consider segregating sites, or nucleotide positions that have undergone a mutation. Second, because a given position can mutate only once, it is sufficient to represent sequences as binary strings, where a 0 indicates the unmutated state and 1 the mutated state. Third, for a given position we can arbitrarily assign the unmutated and mutated states. The infinite-sites model is considered a reasonably good model for long genomes.

The four-gamete test identifies reticulate evolution by looking at pairs of segregating sites. Given biallelic data, there are four possible haplotype patterns, or states, for a pair of segregating sites: 00, 10, 01, or 11.[19] The statement of the four gamete test is this: in any given dataset, the simultaneous presence of all four haplotype states in any pair of segregating sites is incompatible with strictly clonal evolution, and indicates reticulate evolution. To see this, assume state 00 as the ancestor to states 10 and 01, which arise from two independent mutations. Because of the no multiple-mutation assumption, it is not possible for either of these two states to then independently mutate into state 11. The only

[19]These sites need not be adjacent.

way for state 11 to arise is via a reticulate event that brings together the left site from state 10 and the right site from 01.[20] This process is illustrated in Figure 2.20A.

Under a Hamming metric, the distance matrix for the set of four sequences $s_1 = 00$, $s_2 = 10$, $s_3 = 01$, and $s_4 = 11$ is

$$d = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 2 \end{pmatrix} \tag{2.15}$$

The Vietoris-Rips filtration of this space is shown in Figure 2.20C. At $\epsilon = 0$ the four sequences are disconnected. At $\epsilon = 1$, four edges are drawn, forming a loop. At $\epsilon = 2$, the space is completely connected and the loop is killed. Persistent homology captures the presence of this loop as an H_1 feature in the interval $[1, 2)$ (Figure 2.20D). In this way, the reticulate event is associated with the presence of a nonzero H_1 bar.

We consider this example to be the minimal, or fundamental, unit of reticulation. All more complicated patterns of reticulation can be seen as extensions of this example.

2.3.3 A Complete Example

We illustrate a complete example of how TDA can capture reticulate evolution from complex population data in Figure 2.21. Consider the reticulate phylogeny (Figure 2.21A): five genetic sequences sampled today (yellow circles) originate from a single common ancestor due to clonal evolution (solid blue lines tracing parent to offspring) and reticulate evolution (dotted red lines). In Figure 2.21B, these five samples are placed in the context of a larger dataset, where the data has been projected onto the plane using PCA. Persistent homology is then applied to this larger sample. In Figure 2.21C we demonstrate the construction of a filtered simplicial complex, showing how the connectivity changes as the scale parameter ϵ is increased. Finally, in Figure 2.21D we see the resulting barcode diagram. Using H_0 we can

[20] It is entirely possible for the reticulate event to have had a reversed pattern of ancestry, in which case the reticulation would result in a state 00 and would not be detectable from the sequence data.

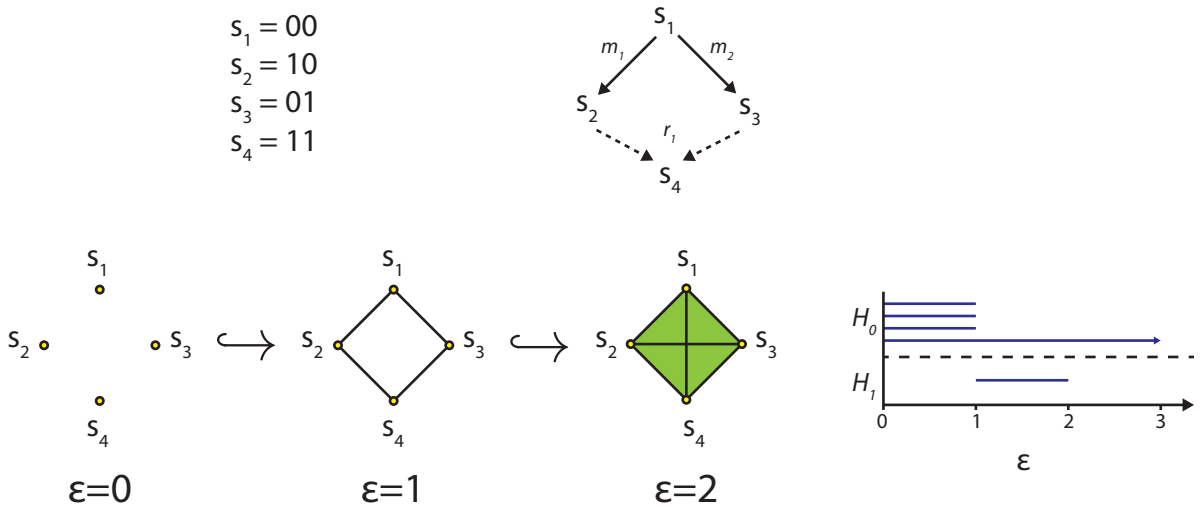


Figure 2.20: The Fundamental Unit of Reticulation. (A) A set of four sequences. (B) An evolutionary genealogy including two mutations (m_1 and m_2) and a single reticulation (r_1). (C) Vietoris-Rips Filtration. (D) Barcode Diagram.

track the number of strains or subclades that persist, roughly corresponding to the tree-like component of the data. The H_1 bar near spanning roughly $\epsilon = 0.13$ to $\epsilon = 0.16$ identifies the presence of a reticulate event. involving the five highlighted sequences. The scale over which this bar persists represents the amount of evolutionary time separating the parents and the reticulate offspring. Additionally, the persistence algorithm will return a generating basis for a particular homology group, which we can use to identify the particular mixtures of sequences involved a reticulation. In this way, we can analyze both the scale and frequency of reticulation in genomic data sets.

We summarize the connection between genomic data and TDA in Table 2.2.

2.3.4 The Space of Trees, Revisited

In Section 2.1.4.4, tree space was introduced as an abstract construction to systematically represent the set of all possible binary trees as a geometric space. Tree space on n leaves, $\mathcal{T}_n$, was shown to be the subspace of the complete space of finite metrics on $\mathbb{R}^{\binom{n}{2}}$ consisting of those metrics that satisfy the four-point condition (or additivity). Because real sequence

Table 2.2: Dictionary connecting algebraic topology and evolutionary biology

Algebraic Topology	Evolutionary Biology
Filtration value ϵ	Genetic distance (evolutionary scale)
0-dimensional Betti number at filtration value ϵ	Number of clusters at scale ϵ
Generators of 0-D homology	A representative element of the cluster
Hierarchical relationship among generators of 0-D homology	Hierarchical clustering
1-D Betti number	Lower bound on number of reticulate events
Generators of 1-D Homology	Reticulate events
Generators of 2-D Homology	Complex horizontal genomic exchange
Non-zero high-dimensional homology (topological obstruction to phylogeny)	No treelike phylogenetic representation exists
Number of higher-dimensional generators over a time interval (irreducible cycle rate)	Lower bound on recombination/reassortment rate

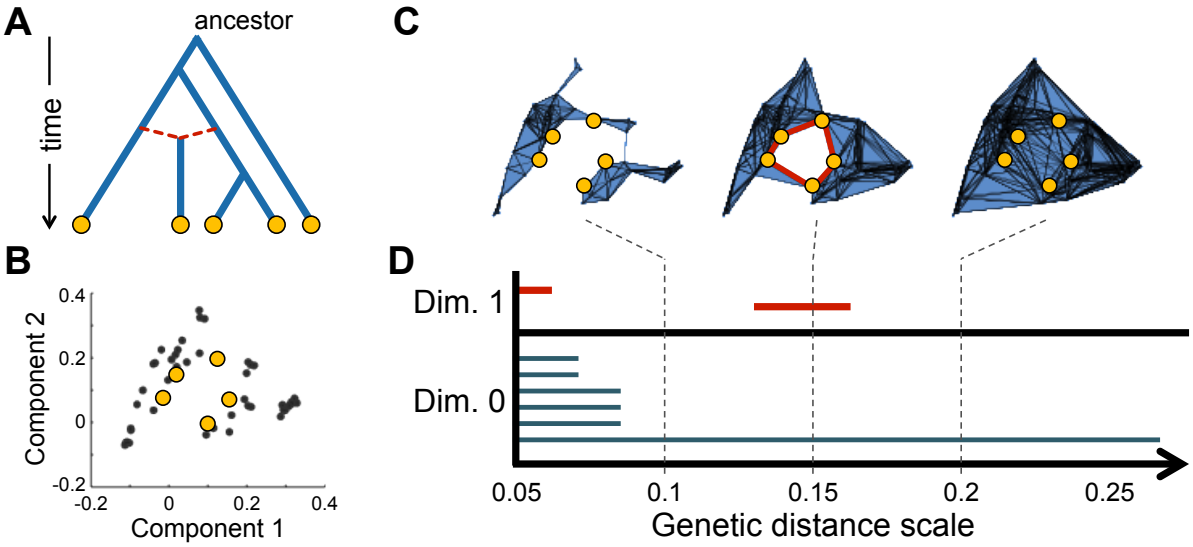


Figure 2.21: Applying persistent homology to genomic data. (A) An evolutionary genealogy including reticulation. (B) Data projected into 2-dimensions. (C) Construction of a filtered simplicial complex. (D) The resulting multiscale barcode diagram.

data will very rarely satisfy this condition, one possible interpretation of phylogenetics is of finding the best projection onto tree space for arbitrary data.

The program we propose can be understood as an extension of the tree space framework. We would like to use topological invariants, specifically homology, to measure the frequency and scale of reticulate evolution in sequence data. From the theorem due to Carlsson, we know that higher homology will vanish for on tree space. And from the simple example in Section 2.3.2, we know that non-additive reticulate processes will have nonvanishing higher homology. Rather than attempt to characterize arbitrary data by projecting onto tree space, we will use persistent homology to compute homological invariants that will characterize our space. Our hypothesis, then, is that as the data moves further from tree space, the less additive it will be, and the stronger from higher homology will be. Our updated picture is shown in Figure 2.22, where we depict tree space embedded in the larger space of finite metrics. We anticipate for appreciable datasets, our sensitivity will be such that those close to the tree space will have little to no homological signal, while those further will have an increasing homological signal. A second point to make is that the metric structure will now allow us to pass through regions of space that are nonadditive. Hence, one could conceivably compare two trees by drawing the direct path between them in metric space, and evaluating the persistent homology at each point.

We note that at the outset of this work, an ambitious goal was set to provide a complete geometric characterization of the space of finite metrics in terms of their topological invariants as measured by persistent homology. While some interesting work has explored the combinatorial structure of the space of metrics on low numbers of points (see [127]), it does not appear feasible in general to provide a complete decomposition.

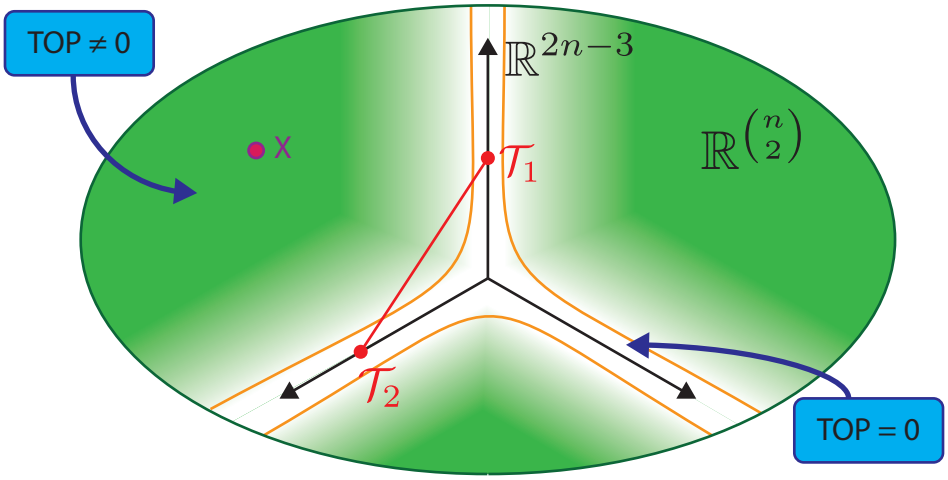


Figure 2.22: Tree space, $\mathcal{T}_n$ is a subspace of the larger space of the metric cone on $\mathbb{R}^{\binom{n}{2}}$. In the presence of reticulate evolution, data may not sit near an additive tree, as for example data X (pink circle). The invariant TOP (topological obstruction to phylogeny) is one way to characterize these spaces and can be computed using persistent homology. In theory, as one moves further from tree space TOP should increase. We study this in more detail in Chapter XX. The orange lines indicate regions close to tree space in which TOP is insensitive to non-additivity. In this example we also indicate two trees $\mathcal{T}_1$ and $\mathcal{T}_2$ which sit on subspaces of different topology in tree space. The direct path between the two topologies will pass through a TOP$\neq$0 region.

Part I

Theory

Chapter 3

Quantifying Reticulation Using Topological Complex Constructions

3.1 Introduction

In Chapter 2, phylogenetic networks were introduced as a generalization of phylogenetic trees as a way of representing reticulation in an evolutionary dataset. In this chapter we introduce additional constructions to extract reticulation information from sequence datasets with higher sensitivity than Vietoris-Rips. We also introduce a method of computing Čech comeplexes on binary sequence data.

The application of persistent homology to molecular sequence data was introduced in [29], where recombination rates in viral populations were estimated by computing L_p norms on barcode diagrams. In that paper, it was shown that persistent homology provides an intuitive quantification of reticulate evolution in sequence data by measuring deviations from tree-like additivity. Reticulation is manifest as nonvanishing higher homology ($H_n > 0$ for $n > 0$) in the filtration. Using persistent homology asa tool to measure reticulate evolution is useful because it

(1) provides a method of quantifying the extent of reticulation, and (2) provides a method

of tracking the scale of reticulate events.

Our goal is to more clearly understand the topological signal that persistent homology captures when applied to sequence data. In doing so, we construct simple examples in which a genetic distance filtration is insensitive to reticulation.

Due to the coarseness of the distance filtration, only those reticulations which have sufficiently strong support in the sequence data will be detected. By coarseness, we mean that thae distance filtration... Small distortions in the metric space, due possibly to incomplete population sampling or weakly supported reticulations, will reduce sensitivity. Looking to increase the resolution of our approach led us to consider a class models which construct a *median graph* from a set of sequences. Median graphs form the basis for a large number of phylogenetic network algorithms and have been extensively studied over the past several decades [9]. The approach is closely related to split decomposition, and it can be shown that the objects resulting from the two methods are identical [8]. The median graph approach imputes putative evolutionry ancestors into the set of vertices, and forms a network representing the incompatibile splits present in the sequence data. A common task has been to quantify the complexity of the resulting network. We show that a filtration of complexes built from the median graph vertex set is a fast an efficient way to characterize the complexity of a phylogenetic network. Due to a result of Gromov, we know that the complexes built on this vertex set will be cubical, making the barcode diagram simple to interpret Gromov [68].

Additionally, we sought to more clearly interpret nontrivial higher homology (H_n for $n > 1$) in the barcode diagram. In Chan, Carlsson, and Rabadan [29], higher homology was presented as evidence for complex reticulations. An application to the 2013 H7N9 influenza epidemic was presented, where the source of the epidemic was shown to be the result of a triple reassortment from three parental strains. The triple reassortment was We expand on this idea, identifying conditions for which higher homology will be observed. These conditions take the form of analogues of the classical four-gamete test. Relationships

between the homology dimension and the number of haplotypes are suggested. To simplify the intepretation of higher homology, we introduce a new construction for building Čech complexes on binary sequence data.

In this paper we present three ideas to increase the usefulness of the signal generated by persistent homology.

The structure of this paper is as follows. In Section 3.2 we review the application of persistent homology to sequence data. We present simple examples in which the genetic distance filtration fails to capture reticulation. In Section 3.5 we present the median closure of the original vertex set. We show how this operation recovers invariant signals of incompatibility in a quantitative way. In Section 3.6 we discuss interpretations of higher dimensional homology and introduce a Čech complex construction on sequence data. In Section 3.4 we present examples of our approach. Throughout, we assume biallelic data under an infinite sites model with no back mutation.

3.2 Persistent Homology of Sequence Data

In this section we briefly review the ideas in Chan, Carlsson, and Rabadan [29] as they relate to the application of persistent homology to sequence data.

3.2.1 Vertical Evolution

In the standard model of evolution, novel genotypes arise via mutation during reproduction. In this case, evolutionary relationships will be accurately modeled as a bifurcating tree. The distance matrix generated from such sequence data will have the property that it is additive. An additive metric can be written as a bifurcating tree such that the distance between any two points in the metric is equal to the path distance along the tree.

To check that a given metric is additive, it is sufficient to check the *four point condition*. The four point condition says that for every set of four points in the data, there is an ordering

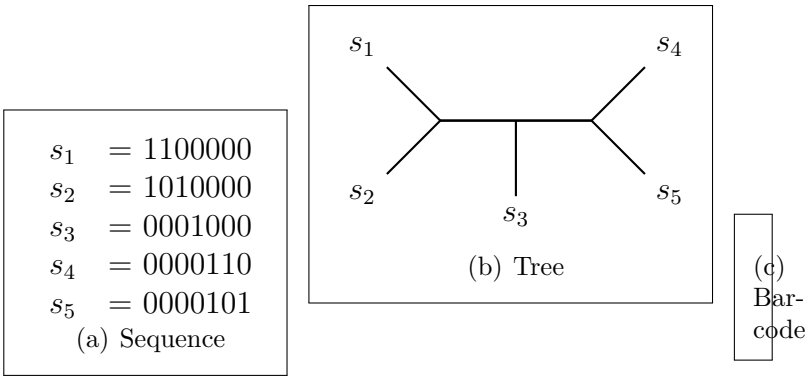


Figure 3.1: A tree is trivially contractible and has vanishing higher homology.

on the points such that

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}. \tag{3.1}$$

Consider the example in Figure 3.1(a). This set of five sequences can be represented the tree in Figure 3.1(b). The barcode diagram from a persistent homology computation is shown in Figure 3.1(c).

A tree is trivially contractible, and hence has vanishing higher homology. This result was proven for sequence data in [29]. In practice, most data is not additive. The field of phylogenetics is essentially tasked with finding the *best* tree given sequence data, for some notion of best.

3.2.2 Reticulate Evolution

Reticulate, or horizontal, evolution refers to any evolutionary process by which genetic material is transfered between organisms in a method other than asexual reproduction. Examples include species hybridization, bacterial gene transfer, and homologous recombination. In these situations, no tree can be drawn that accurately reflects the evolutionary history of a set of sequences.

A simple test for the presence of reticulation is given by the *four gamete test*. The four gamete test states that the simultaneous presence of haplotype patterns 00, 01, 10, and 11

is incompatible with strictly vertical evolution in an infinite sites model. It provides direct evidence for reticulate evolution. One way to quantify recombination in a set of sequences is the Hudson-Kaplan test, which counts the minimum number partitions required in the data such that within each partition the all sites are compatible [73].

We consider the four gametes to be the fundamental unit of recombination. Topologically, this unit represents a loop. In a persistent homology computation, we would see nonvanishing H_1 homology in the interval $[1, 2)$ (see Figure XXX).

In the fundamental loop, we can give an interpretation to each vertex. There is a common ancestor, two parents, and a recombinant child. Of course, we do not *a priori* know which sequences played which role in a given loop, which is the same as the problem of rooting a phylogenetic tree. Persistent homology is simply a method of efficiently counting the number of such loops in the data, across all genetic scales.

3.3 Reticulation Quantification Using Homology

We can use persistent homology to quantify reticulation in a particular dataset. There are some flaws in the original Vietoris-Rips construction. This approach generalizes that construction in order to recover additive trees.

3.4 Examples

In considering small examples of this form we often encountered cases in which the four gamete test indicated reticulation, but persistent homology failed to detect a loop. What these examples had in common is that due to distortion in the metric space, simplices would collapse before they should have. This could have been due to incomplete population This could have been due to incomplete sampling, in which case recombination fails to be detected because parental sequences collapse to early, or due to cases where recombination creates

new sequences that sit intermediate to parental and ancestral sequences. Here we work through two examples in detail.

Example 1 It is generally the case that we do not have a complete sampling of the sequences corresponding to the evolutionary history of a set of sequences. For example, we may not have sampled the true recombinant child, only a descendant which has accumulated additional mutations. Consider the set of sequences 000, 100, 010, and 111. From the four-gamete test we know there is an incompatibility between sites 1 and 2, indicating the presence of a reticulate event. Let us arbitrarily choose s_1 to be the common ancestor, s_2 and s_3 to be parents, and s_4 to be a descendant of the reticulate event. We can infer that the recombinant was of the form $s_r = 110$. Unfortunately, the persistent homology the four sequences will be trivial. To understand why, consider an embedding of the four sampled sequences onto the 3-cube, as seen in Figure XXX.

The failure to detect the loop is due to the ancestral and parent sequences collapsing before connecting with the recombinant child. In general, for a loop to be detected, the two internal distances must be greater than any of the four side distances. In this case, the internal distance from parent 1 (s_2) to parent 2 (s_3), d_{23} is equal to the distances from each parent to the sampled descendent of the recombinant (d_{24} and d_{34}). This is a general issue with the application of persistent homology to phylogenetic data. Distortions in the metric space due to incomplete sampling can lower the detection sensitivity, even in cases where incompatible sites are present. In this example, had we sampled the recombinant child (white vertex), persistent homology would detect the loop between s_1, s_2, s_3, and s_r. s_4 would be seen as the descendant of s_r. In the following section we will introduce a method of imputing missing points into the vertex set using the median closure operation. The result will be an augmented simplicial complex, formed from a new vertex set consisting of the original data and points added from the median operation, which we call the *median complex*.

Example 2 This example is taken from Song and Hein [126]. Consider the set of sequences: $s_1 = 0000$, $s_2 = 1100$, $s_3 = 0011$, $s_4 = 1010$, and $s_5 = 1111$. There are pairwise incompatibilities between sites 1 and 3, 1 and 4, 2 and 3, and 2 and 4. Performing the Hudson-Kaplan test yields $R_M = 1$, with a partition between sites 2 and 3. Song and Hein [126] showed that a minimum of two recombinations were required to explain this data. In this example, persistent homology will contract immediately, with trivial higher homology. To understand why this is the case, consider an embedding into $\mathbb{R}^3$. The problem is that s_3 sits in the middle of the other four sequences, and at $\epsilon = 2$ everything contracts. Had s_3 not been present in the data, we would have had an example very similar to Example 3.4, with the interpretation of one recombination event. We term this the "dixie cup" example. The conclusion to draw from this example is that multiple recombination events can interact in complicated ways, destroying signal from persistent homology.

3.5 The Median Complex Construction

Definition 1. For any three aligned sequences a, b, and c, the *median* sequence $m(a, b, c)$ is defined such that each position of the median is the majority consensus of the three sequences.

For example, consider the three sequences $a = 110$, $b = 011$, and $c = 101$. At each site we have the set $\{1, 1, 0\}$. The majority consensus for each site is 1, therefore the median sequence is $m = 111$. In any further analysis, we augment the original data to include the computed median sequence. Note that as defined here, the median operation is defined only for binary sequences.

Having defined the median operation, we now define the *median closure*. Given an alignment S, the median closure, $\bar{S}$, is defined as the vertex set generated from the original set S that is closed under the median operation,

$$\bar{S} = \{v \colon v = m(a, b, c) \in \bar{S} \forall a, b, c \in \bar{S}\} \tag{3.2}$$

We can obtain the median closure $\bar{S}$ by repeatedly applying the median operation to sets of three sequences until no new sequences are added. Effectively, computing the median closure imputes interior nodes into the dataset. We call complexes formed from the original sequences the *leaf complexes*, and call complexes formed from the median closure the *median complexes*. We can then proceed by computing the persistent homology of this median closure. The downside of the median closure operation is that we can no longer identify the loops we measure as reticulate events. The median closure operation can generate multiple loops from a single incompatibility. Let us now reconsider our two examples from the previous section, under the median closure.

Example 1 We add one median vertex, $m(s_2, s_3, s_4) = 110$ (Figure XX). Persistent homology now detects an H_1 interval in the range $\epsilon = [1, 2)$.

Example 2 We add four median vertices (Figure XX). Persistent homology detects four H_1 intervals in the range $\epsilon = [1, 2)$.

Filtrations on Buneman graphs have been defined previously [47], but not using an explicit sequence representation. The filtration defined in Dress, Huber, and Moulton [47] is based on a complicated polytope construction scheme defined directly from the split decomposition. Given that all median graphs are split networks [74], the constructions are identical but the extracted information is not. To the best of our knowledge, quantification of the complexity of these objects has not been measured using homological tools.

3.5.1 Inclusion

We have examined the persistent homology of two topological constructions on sequence data: the leaf complex and the median complex. Counting β_1 intervals in the leaf complex underestimates reticulate evolution because of incomplete sampling, while counting β_1 intervals in the median complex overestimates reticulate evolution. The median complex is

in some sense an upper bound on probable recombination histories, and contains within it all possible recombination graphs within it (not strictly true, as there are infinitely many complicated ARGs - but it does contain within it all maximum parsimony trees). We can hypothesize that there exists a true complex, called the *evolutionary complex*, which will accurately reflect the evolutionary relationships in the sequences. Information about the evolutionary complex is not available to us, however we can say that there exists an inclusion between the homotopy types of the three complexes

$$\mathrm{Cl}(\mathcal{LC}) \hookrightarrow \mathrm{Cl}(\mathcal{EC}) \hookrightarrow \mathrm{Cl}(\mathcal{MC}) \tag{3.3}$$

Recovery of an optimal $\mathcal{EC}$ is the task of many ARG-based methods and is known to be an NP-hard problem and is not considered here. For example, given an $\mathcal{EC}$ as computed from some other tool, we might be able to say something useful about the topological complexity.

3.5.2 Split Decomposition

Split decomposition can take a distance matrix and reduce it to a set of weighted splits.

3.6 Interpretation of Higher Dimensional Homology

In Chan, Carlsson, and Rabadan [29] it was argued that higher dimensional homology (H_d for $d > 1$) is evidence for 'more complex' reticulate events. Here we try to make this notion more precise, showing by way of examples that higher dimensional homology can be interpreted as evidence of multiple interacting reticulate events. First, we detour slightly and introduce a Čech complex construction that will increase our sensitivity to these events.

3.7 Čech Complex Construction as an Optimization Problem

The Čech complex is defined on a set of points S as

$$\check{\text{C}}\text{ech}(r) = \left\{ \sigma \subseteq S | \bigcap_{x \in \sigma} B_x(r) \neq \emptyset \right\}, \tag{3.4}$$

where $B_x(r)$ is the ball of radius r centered at vertex x. By the nerve lemma, the homotopy type of the Čech covering is guaranteed to be identical to that of the original topological space [17].

Computing the Čech complex is often an expensive operation, such that in practice the Vietoris-Rips complex is used. Unlike the Vietoris-Rips complex, which is entirely defined by the 1-skeleton, the Čech complex requires one to check each simplex σ up to some maximum dimension D. The Čech complex therefore requires one to know the ambient space the data is embedded in, unlike a Rips complex which can be built directly from distance data. Binary sequence data of length d explicitly sits on the discrete lattice of $\{0,1\}^d$ with an L_1 norm. In this case, it is not immediately obvious how to define when three sequences should form a simplex. One Therefore, we expand the ambient space to $\mathbb{R}^d$ with an L_1 metric. This choice of metric is motivated by two reasons. First, the L_1 norm maintains the Hamming distance between sampled points. Second, the L_1 norm keeps the primary theorem intact, that is tree like data generates trivial homology. [1]

The problem of deciding if a particular simplex σ belongs in the Čech complex at radius r is the same as checking if a ball of radius r can be placed such that each point x in σ is contained within the ball. In $\mathbb{R}^d$ with an L_2 metric there exists an efficient randomized algorithm for computing this radius known as the *miniball algorithm.*[62] However, the efficiency of the miniball algorithm relies on the strict convexity of the L_2 metric and therefore is not applicable to a space with an L_1 metric. Instead, we pose the miniball problem in L_1 as a

[1] This notion has a natural extension to multiallelic sites which is not detailed here.

generic convex optimization problem, and use standard library solver. That is, we define a $d+1$ dimensional optimization problem where x is the miniball center and R is the miniball radius.

The problem is stated as

$$\begin{aligned} \text{minimize} \quad & R \\ \text{subject to} \quad & \forall p \in P : ||x - p||_1 \leq R \\ & x \in \mathbb{R}^d \end{aligned}$$

We implement the problem in `cvxpy`. TODO: A brief comment about the complexity of this routine. The randomized miniball algorithm has constant complexity in dimension.

3.7.1 Molecular Hypothesis

Gromov proved that a median graph is the 1-skeleton of a CAT(0) cubical complex [68]. The homology of a cubical complex can be efficiently computed using the methods of Kaczynski, Mischaikow, and Mrozek [80] through a slightly different construction. We define a cubical flag complex and build a filtration dimension by dimension (to expand on this point...) The barcode diagram will then have the natural interpretation of being composed of sets of hypercubes of varying dimension. If we consider each bar of dimension n in the barcode diagram in turn, we can determine the incompatible sites that it represents. Dimension 1 bars (2-cubes) will have one pair of incompatible sites with four haplotypes. Dimension 2 bars (3-cubes) will have three pairs of incompatible sites with eight haplotypes. In general, n bars will represent $n+1$-cubes in which all $2^{(n+1)}$ haplotypes are present in the vertices of the generating cycle.

From the barcode diagram it will not in general be possible to decompose our construction into the primitive building blocks of hypercubes. This is because the hypercubes of dimension ($n > 2$) will in general not be independent, but can interact by sharing lower dimensional faces. Nonetheless, to aid in decomposing the barcode diagram, we constructed the following

table, which contains the homology ranks (betti numbers) for powers of the hypercube graph, computed using the Čech complex. Incidentally, it was understanding the structure of numbers in a table very much like Table 3.7.1 which led us to find a method of computing Cech homology instead of Rips homology.

$d =$	1	2	3	4	5	6
H_0	2	4	8	16	32	64
H_1	0	1	5	17	49	129
H_2	0	0	1	7	31	111
H_3	0	0	0	1	9	49
H_4	0	0	0	0	1	11
H_5	0	0	0	0	0	1
H_6	0	0	0	0	0	0

Table 3.1: Čech Homology of Hypercube

We include a simple proof of the numbers in this table [right here].

3.8 Examples

3.8.1 Kreitman Data

A benchmark dataset in recombination studies is the Kretiman data [84]. The dataset consists of eleven sequences (nine unique) of the Adh locus from *Drosophilia melanogaster* collected from various locations, with 43 segregating sites. Several methods have been applied to this data to estimate the minimum number of recombinations present in this data. The Hudson-Kreitman test yields 6, while Song-Hein computed 7. The persistent homology of

the original dataset detected no loops. The median closure expanded the dataset to 46 vertices. Here we have non-trivial homology: 32 dimension-1 intervals and 7 dimension-2 intervals. The barcode plot is shown in Figure ??. Can we use the homology information to make a claim about the minimum recombination graph? Can we set an upper bound on the number of recombination graphs?

3.8.2 Buttercup Data

3.8.3 Additional Examples

See Huson, Rupp, and Scornavacca [74].

3.8.4 Simple Examples

Generation of one dimensional homology requires the presence of four incompatible haplotypes (00, 10, 01, 11). That is, there is a condition on pairs of segregating sites, and at least two sites will be required to generate H_1. Homology of dimension $n > 1$ will be a higher order effect and require the interaction of multiple pairs of sites. One might surmise that all possible haplotypes on n segregating sites are required to generate homology of dimension $n-1$. For example, on the 3-cube, there are eight haplotypes. H_2 is generated in the interval $[1.0, 1.5)$.

In fact, subsets of the 3-cube generating H_2 can be formulated. Consider the set of sequences $S = (000, 100, 010, 001, 111)$. The persistent homology of S will generate H_2 in the interval $[1, 1.5)$. A possible evolutionary scenario is presented in Figure XXX. We see that sequence s_5 is a triple reassortment of sequences s_2, s_3, and s_5. Further, notice that there is total incompatibility between sites $(1, 2)$, $(2, 3)$, and $(1, 3)$. Contrast this with the example detailed in Figure XXX. Here, we have a set of six sequences which exhibits two H_1 loops, and no H_2 homology. The two loops can be seen as independent. And if we examine

3.9 Conclusions

Persistent homology can capture and quantify complex patterns of reticulation in genomic data. The standard Vietoris-Rips filtration is susceptible to reduced sensitivity due to incomplete sampling or interactions between reticulations. Constructing the median closure of the original sequence set increases the topological signal of reticulation. Future work will focus on efficient implementations of constructing this closure.

An interesting additional observation is that the number of recombinations required to explain the fully saturated hypercube is exactly equal to the alternating sum of the homology ranks.

Chapter 4

Parametric Inference using Persistence Diagrams

"I predict a new subject of statistical topology. Rather than count the number of holes, Betti numbers, etc., one will be more interested in the distribution of such objects on noncompact manifolds as one goes out to infinity"

Isadore Singer

4.1 Introduction

Computational topology is emerging as a new approach to data analysis, driven by efficient algorithms for computing topological structure in data. Perhaps the most mature tool is persistent homology, which summarizes multiscale topological information in a two-dimensional persistence diagram (see Figure 4.1 and Section 3.2). Recent work has concentrated on developing the statistical foundations for data analysis using the persistent homology framework [56, 16, 32]. The focus of this work has been estimating the topology of an object from a finite, noisy sample. Doing so requires statistical methods to distinguish topological signal from noise.

Here we consider a different scenario. Many simple stochastic models generate complex

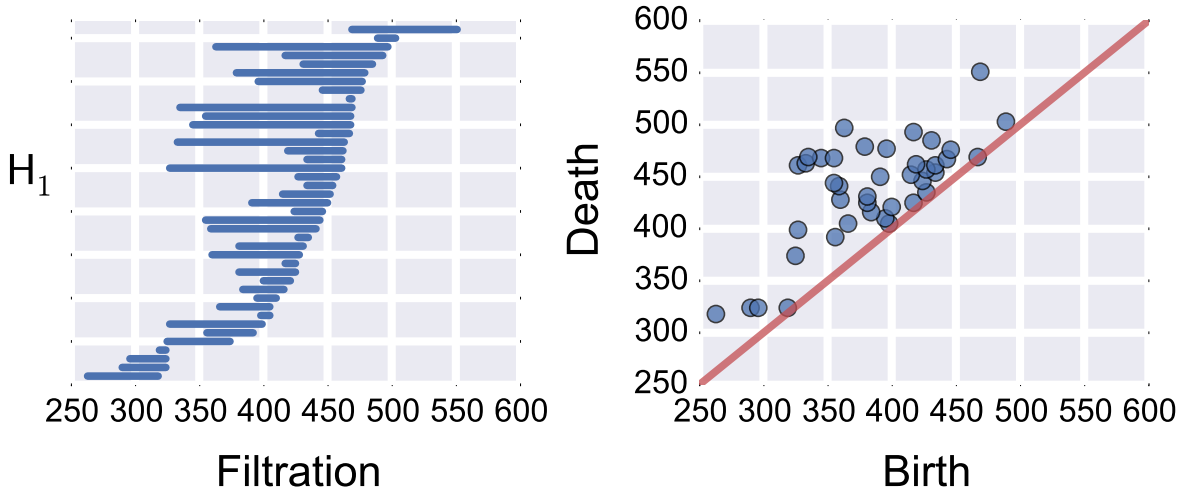


Figure 4.1: Two representations of the same topological invariants, computed using persistent homology. Left: Barcode diagram. Right: Persistence diagram. Data was generated from a coalescent simulation with $n = 100$, $\rho = 72$, and $\theta = 500$.

data that cannot be readily visualized as a manifold or summarized by a small number of topological features. These models will generate persistence diagrams whose complexity increases with the number of sampled points. Nevertheless, the collection of measured topological features may exhibit additional structure, providing useful information about the underlying data generating process. While the persistence diagram is itself a summary of the topological information contained in a sampled point cloud, to perform inference further summarization may be appropriate, e.g. by considering distributions of properties defined on the diagram. In other words, we are less interested in learning the topology of a particular sample, but rather in understanding the expected topological signal of different model parameters.

In this chapter, we show that summary statistics computed on the persistence diagram can be used for likelihood-based parametric inference. We use genomic sequence data as a case study, examining the topological behavior of the coalescent process with recombination, a widely used stochastic model of biological evolution. We find that the process generates

nontrivial topology in a way that depends sensitively on parameter in the model. The idea is presented as a proof of concept, in order to motivate the identification additional models with regular topological structure that may amenable to this type of inference.

4.2 Warmup: Gaussian Random Fields

Here we show that parametric inference of Gaussian Random Fields can be performed from the barcode diagram. Make reference to [1]. Connections with problems in cosmology.

4.3 The Coalescent Process

The coalescent process is a stochastic model that generates the genealogy of individuals sampled from an evolving population [136]. The genealogy is then used to simulate the genetic sequences of the sample. This model is essential to many methods commonly used in population genetics. Starting with a present-day sample of n individuals, each individual's lineage is traced backward in time, towards a mutual common ancestor. Two separate lineages collapse via a coalescence event, representing the sharing of an ancestor by the two lineages. The stochastic process ends when all lineages of all sampled individuals collapse into a single common ancestor. In this process, if the total (diploid) population size N is sufficiently large, then the expected time before a coalescence event, in units of $2N$ generations, is approximately exponentially distributed:

$$P(T_k = t) \approx \binom{k}{2} e^{-\binom{k}{2}t}, \tag{4.1}$$

where T_k is the time that it takes for k individual lineages to collapse into $k-1$ lineages.

After generating a genealogy, the genetic sequences of the sample can be simulated by placing mutations on the individual branches of the lineage. The number of mutations on each branch is Poisson-distributed with mean $\theta t/2$, where t is the branch length and θ is the

population-scaled mutation rate. In this model, the average *genetic distance* between any two sampled individuals, defined by the number of mutations separating them, is θ.

The coalescent with recombination is an extension of this model that allows different genetic loci to have different genealogies. Looking backward in time, recombination is modeled as a splitting event, occurring at a rate determined by population-scaled recombination rate ρ, such that an individual has a different ancestor at different loci. Evolutionary histories are no longer represented by a tree, but rather by an *ancestral recombination graph*. Recombination is the component of the model generating nontrivial topology by introducing deviations from a contractibile tree structure, and is the component which we would like to quantify. Coalescent simulations were performed using `ms` [72].

4.4 Statistical Model

The persistence diagram from a typical coalescent simulation is shown in Figure 4.1. Examining the diagram, it would be difficult to classify the observed features into signal and noise. Instead, we use the information in the diagram to construct a statistical model in order to infer the parameters, θ and ρ, which generated the data. Note that we consider inference using only H_1 invariants, but the ideas easily generalize to higher dimensions. We consider the following properties of the persistence diagram: the total number of features, K; the set of birth times, $(b_1, \ldots, b_K)$; the set of death times, $(d_1, \ldots, d_K)$; and the set of persistence lengths, $(l_1, \ldots, l_K)$. In Figure 4.2 we show the distributions of these properties for four values of ρ, keeping fixed $n = 100$ and $\theta = 500$. Several observations are immediately apparent. First, the topological signal is remarkably stable. Second, higher ρ increases the number of features, consistent with the intuition that recombination generates nontrivial topology in the model. Third, the mean values of the birth and death time distributions are only weakly dependent on ρ and are slightly smaller than θ, suggesting that θ defines a natural scale in the topological space. However, higher ρ tightens the variance of the distributions. Finally,

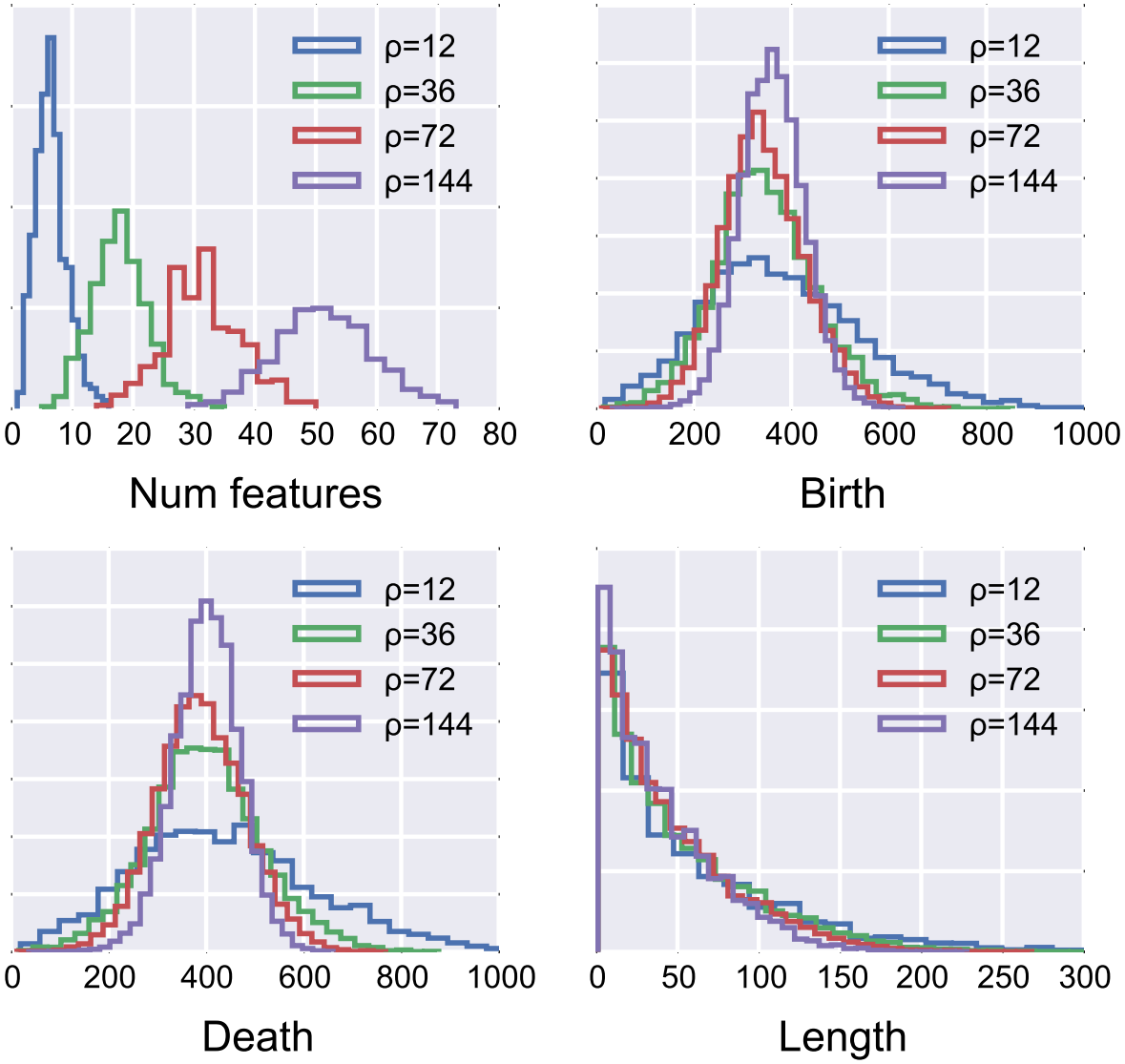


Figure 4.2: Distributions of statistics defined on the H_1 persistence diagram for different model parameters. Top left: Number of features. Top right: Birth time distribution. Bottom left: Death time distribution. Bottom right: Feature length distribution. Data generated from 1000 coalescent simulations with $n = 100$, $\theta = 500$, and variable ρ.

persistence lengths are independent of ρ.

Examining Figure 4.2, we can postulate: $K \sim \text{Pois}(\zeta)$, $b_k \sim \text{Gamma}(\alpha, \xi)$, and $l_k \sim \text{Exp}(\eta)$. Death time is given by $d_k = b_k + l_k$, which is incomplete Gamma distributed. The parameters of each distribution are assumed to be an *a priori* unknown function of the model parameters, θ and ρ, and the sample size, n. Keeping n fixed, and assuming each element in the diagram is independent, we can define the full likelihood as

$$p(D \mid \theta, \rho) = p(K \mid \theta, \rho) \prod_{k=1}^{K} p(b_k \mid \theta, \rho) p(l_k \mid \theta, \rho). \tag{4.2}$$

Simulations over a range of parameter values suggest the following functional forms for the parameters of each distribution. The number of features is Poisson distributed with expected value

$$\zeta = a_0 \log\left(1 + \frac{\rho}{a_1 + a_2 \rho}\right) \tag{4.3}$$

Birth times are Gamma distributed with shape parameter

$$\alpha = b_0 \rho + b_1 \tag{4.4}$$

and scale parameter

$$\xi = \frac{1}{\alpha}(c_0 \exp(-c_1 \rho) + c_2). \tag{4.5}$$

These expressions appears to hold well in the regime $\rho < \theta$, but break down for large ρ. The length distribution is exponentially distributed with shape parameter proportional to mutation rate, $\eta = \alpha\theta$. The coefficients in each of these functions are calibrated using simulations, and could be improved with further analysis. This model has a simple structure and standard maximum likelihood approaches can be used to find optimal values of θ and ρ.

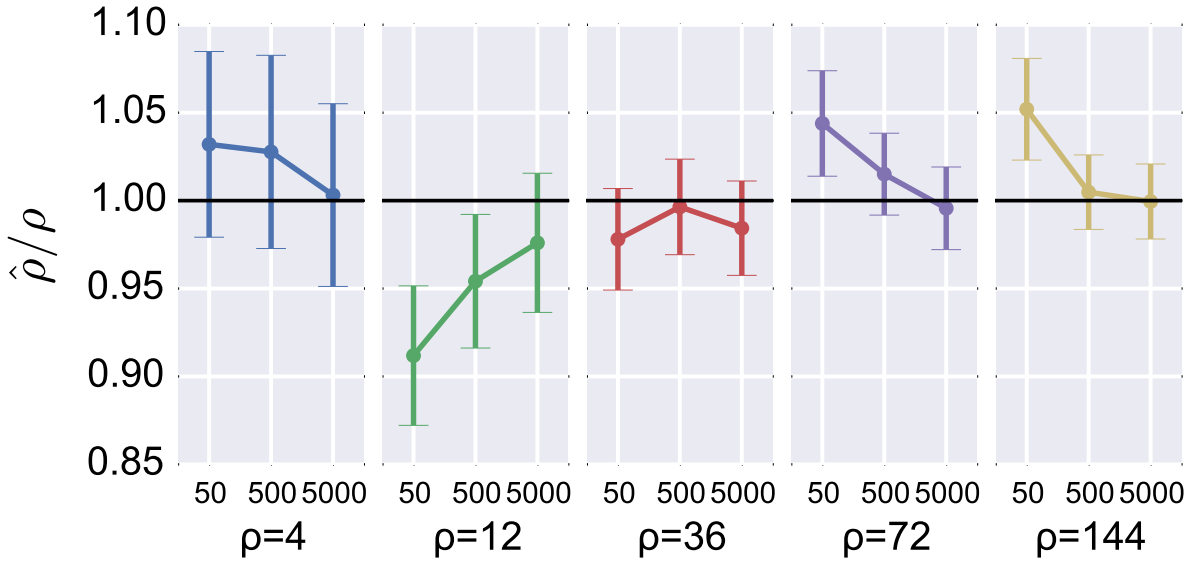


Figure 4.3: Inference of recombination rate ρ using topological information. The recombination rate ρ is estimated for five values $\{4, 12, 36, 72, 144\}$ at three different mutation rates $\{50, 500, 5000\}$. Mean estimate over 500 simulations and 95% confidence interval is shown.

4.5 Experiments

4.5.1 Coalescent Simulations

We simulated a coalescent process with sample size $n = 100$ and $l = 10{,}000$ loci. The mutation rate, θ, was varied across $\theta = \{50, 500, 5000\}$. The recombination rate, ρ, was varied across $\rho = \{4, 12, 36, 72\}$. The output of the process is a set of binary sequences of variable length (length is dependent on θ). The Hamming metric is used to construct a pairwise distance matrix between sequences. We computed persistent homology and used the model described in Section 4.4 to estimate θ and ρ. Results are shown in Figure 4.3, where we plot estimates and 95% confidence interval from 500 simulations. We observe an improved ρ estimate at higher mutation rate. This is expected, as increasing θ is essentially increasing sampling on branches in the genealogy. We also observe tighter confidence intervals at higher recombination rates, consistent with the behavior seen in Figure 4.2.

4.6 Conclusions

In machine learning, the task is often to infer parameters of a model from observations. In this chapter we have presented a proof of concept for statistical inference based on topological information computed using persistent homology. Unlike previous work, which considered estimating homology of a partially observed object, we were interested in a model which generates a complex, but stable, topological signal. Three conditions were required for the success of this approach: First, a well-defined statistical model. Second, an intuition that the observed topological structure is directly correlated with the parameters of interest in the model. Third, sufficient topological signal to reliably estimate statistics on the persistence diagram. It is an open question to identify classes of models for which these conditions will hold.

Part II

Applications

Chapter 5

Phage Mosaicism

5.1 Introduction

Phages are microbial viruses which can infect bacteria, archaea, or single-celled eukaryotes. By some measures, they are the most abundant and diverse class of organism on the planet. It is estimated that there are 10^{31} extant bacteriophages [120].[1] The phage population completely turns over every few days – an estimated infection rate of 10^{23} per second [128].

Phages play an essential role in natural ecosystems by regulating bacterial populations. Steps have been taken towards harnessing this ability for productive use – the FDA has approved several bacteriophage products designed to kill harmful bacteria in dairy and meat products [20]. Also promising are potential phage therapies for treating pathogenic bacterial infections, although research in this direction is controversial [81].

Phages are classified based on lifestyle: virulent phages have a lytic life cycle and will infect a host, multiply, and exit the cell via lysis, killing the host organism; temperate phages have a lysogenic life cycle and can remain within the host in a latent state, without disrupting host cellular function. Phages can have a nucleic acid composition that is either double-

[1] The estimate can be arrived at two independent ways: by assuming a total bacterial population size of 10^{30}, and approximately ten phages per bacteria; or by the observation of a phage density of 10^6 to 10^7 per mL of seawater.

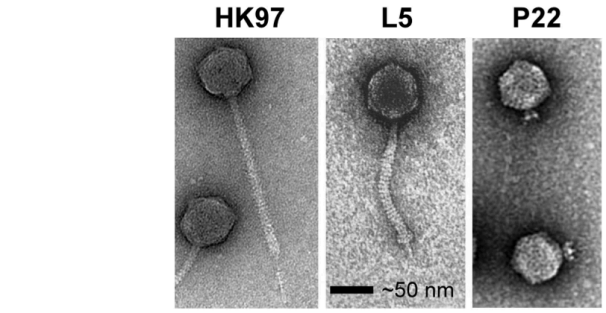
Table 5.1: Phage families defined by the ICTV

Order	Family	Morphology	Nucleic acid
Caudovirales	*Myoviridae*	Nonenveloped, contractile tail	linear dsDNA
	Siphoviridae	Nonenveloped, noncontractile tail (long)	linear dsDNA
	Podoviridae	Nonenveloped, noncontractile tail (short)	linear dsDNA
Ligamenvirales	*Lipothrixviridae*	Enveloped, rod-shaped	linear dsDNA
	Rudiviridae	Nonenveloped, rod-shaped	linear dsDNA
Unassigned	*Ampullaviridae*	Enveloped, bottle-shaped	linear dsDNA
	Bicaudaviridae	Nonenveloped, lemon-shaped	circular dsDNA
	Clavaviridae	Nonenveloped, rod-shaped	circular dsDNA
	Corticoviridae	Nonenveloped, isometric	circular dsDNA
	Cystoviridae	Enveloped, spherical	segmented dsRNA
	Fuselloviridae	Nonenveloped, lemon-shaped	circular dsDNA
	Globuloviridae	Enveloped, isometric	linear dsDNA
	Guttaviridae	Nonenveloped, ovoid	circular dsDNA
	Inoviridae	Nonenveloped, filamentous	circular ssDNA
	Leviviridae	Nonenveloped, isometric	linear ssRNA
	Microviridae	Nonenveloped, isometric	circular ssDNA
	Plasmaviridae	Enveloped, pleomorph	circular dsDNA
	Tectiviridae	Nonenveloped, isometric	linear dsDNA

stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), or single-stranded RNA (ssRNA). Of these, dsDNA is by far the most common. The typical phage genome length is on the order of 10^5 bases, but can range from 10^3 to 10^6 bases.

Because there is no conserved gene across all phage populations, there is no accepted way of constructing a molecular phage taxonomy. The current bacteriophage taxonomy is compiled by the International Committee on Taxonomy of Viruses (ICTV) and is based on virus morphology, host range, lifestyle, and nucleic acid composition [76]. Table 5.1 presents an overview of phage families as defined by the ICTV. There are two assigned orders and eighteen recognized families. Fourteen families have dsDNA, two families have ssDNA, and two families have an RNA genome.

Phages have been shown to be subject to high rates of reticulate genomic exchange [141]. The phage genome was believed to be mosaic, composed of distinct modules that can be freely exchanged within a population. Increased genomic data has confirmed this mosaic structure and raised questions about the applicability and interpretation of the ICTV taxonomy. Based solely on morphology and host, the ICTV taxonomy has been shown to be inconsistent with the genomic data, as the following example from Lawrence *et al.* shows [89]. In Figure 5.1



		HK97	L5	P22
	Viral Family	Siphoviridae	Siphoviridae	Podoviridae
	Bacterial Host	*Escherichia coli*	*Mycobacterium smegmatis*	*Salmonella enterica*
Shared Genes	HK97	-	0 %	20 %
	L5	0 %	-	0 %
	P22	22 %	0 %	-

Figure 5.1: Inconsistency of morphological classifications in bacteriophage. HK97 and L5 are classified in the Siphoviridae family of long tail non-contractile phages, despite sharing no gene content. P22, a short-tail phage in the Podoviridae family, while morphologically dissimilar, shares 20% gene content with HK97. Figure adapted from [89].

we show three different bacteriophage species: Enterobacteria phage HK97, Mycobacterium phage L5, and Enterobacteria phage P22. HK97 is a Siphoviridae infecting *E. coli*. L5 is a Siphoviridae infecting *M. smegmatis*. P22 is a Podoviridae infecting *S. enterica*. HK97 and L5 belong to the Siphoviridae family comprised of long tail noncontractile phages. P22 belongs to the Podoviridae family comprised of short tail phages. Visually, it appears that HK97 and L5 should indeed be classified as distinct from P22. However, genomic analysis reveals that HK97 and L5 share no gene content, and, despite appearances to the contrary, HK97 and P22 share 20% gene content. This example demonstrates that morphology and host range alone are not sufficient in representing phage relationships.

Alternative representations of phage relationships have been proposed based on whole genome analysis. For example, Rohwer and Edwards constructed a phage phylogenetic tree using differences in phage proteomes [119]. Proux *et al.* proposed a phylogenetic representation based on comparative analysis of head and tail sequences [115]. However, these models still make the assumption of tree-like relationships, which will not be appropriate for representing highly mosaic molecular relationships.

In this chapter, we use approaches from topological data analysis to identify, measure, and represent reticulate evolution in a population of phage sequences. This work is primarily based on data collected by Lima-Mendez *et al.* [93]. First, we use persistent homology to characterize reticulation in phage genomes. We find H_0 is largely inconsistent with existing phage taxonomies, and interpret H_1 as evidence for reticulate genetic exchange due to shared ecology and host range. Second, we visualize phage molecular relationships using Mapper, identifying clusters of phages with common gene content and host range. Representative protein families for each phage cluster are identified. The Mapper network suggests an alternate way of representing phage molecular relationships.

5.2 Data

We use data initially collected an analyzed in [93]. The initial data set consists of a collection of 306 sequenced bacteriophage genomes. We show summary information about the data in Figure 5.2. Of the 306 genomes, 246 consist of dsDNA, 36 ssDNA, 12 dsRNA, and 8 ssRNA. Four have unclassified nucleic acid material. With respect to lifestyle, 146 are temperate and 72 are virulent. Actinoplanes phage phiAsp2 is the single pseudotemperate phage, which means it largely maintains a temperate lifestyle but can occasionally enter a virulent state. For 87 phages the lifestyle is unknown. Taxonomically, the vast majority belong to order Caudovirales (221), which comprises Siphoviridae (117), Myoviridae (47), and Podoviridae (54). Order Ligamenviralies (4) comprises Lipothrixviriae (2) and Rudiviridae (2). Unassigned families include Inoviridae (22), Cystoviridae (12), Gokushoviridae (8), and Microviridae (6).

Each of the 306 bacteriophage genomes has been sequenced and annotated.[2] This step resulted in 19,537 unique bacteriophage phage genes. In the original study [93], these genes

[2]The annotation step assigns genes to subsequences of the genome. For well-characterized species this is facilitated by a reference genome. For less well-characterized species this can require the use of heuristic gene-finding algorithms.

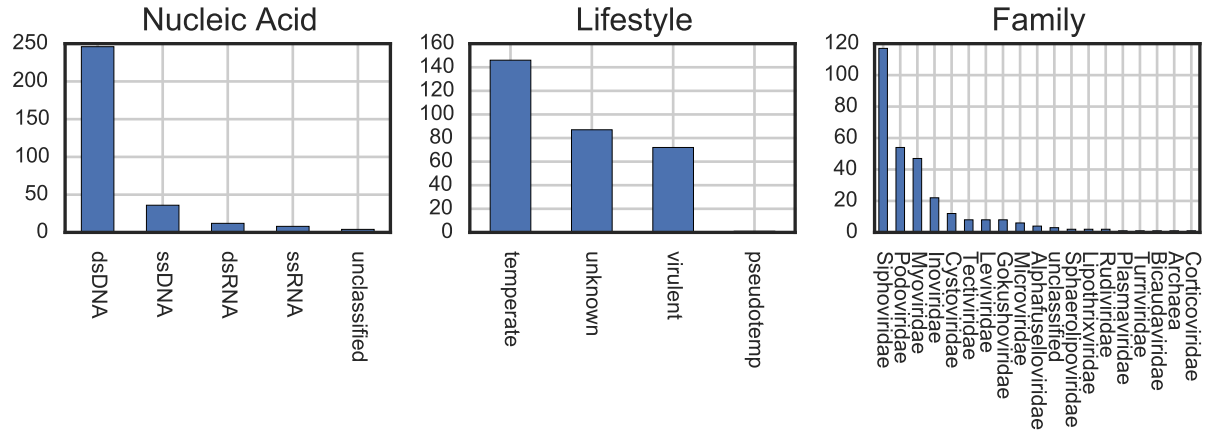


Figure 5.2: Summary annotations of phage data used in this analysis. 306 bacteriophage genomes were included, as originally collected in [93]. Here we show various annotations for the phage, including nucleic acid type, lifestyle, and taxonomic family (as defined by the ICTV). For some phage strains this data is unknown.

were then clustered into 8,576 protein families using BlastP, which analyzes pairwise similarity of proteins [3]. Protein families share homology, which implies some degree of shared evolutionary ancestry. Phages can then be represented as phyletic profiles in a protein family-space, indicating the presence or absence of a particular protein family. In this case, the phyletic matrix P is a 306×8576 binary matrix.

5.3 Measuring Phage Mosaicism with Persistent Homology

We apply persistent homology to the phyletic profiles in order to quantify reticulation in the bacteriphage data. Because we have transformed from sequence space into phyletic profiles, we do not invoke a specific evolutionary model. However, the fundamental theorem that non-trivial homology implies reticulation still holds. First, we construct an appropriate metric space. Following [93], we use a hypergeometric model as follows. For two phages A and B, let a be the number of protein families in phage A, b be the number of protein

families in phage B, and c be the number of protein families in common. Let n be the total number of protein families. Then we can compute the p-value that the number of shared protein families c is significant as

$$P_{AB} = \sum_{i=c}^{\min(a,b)} \frac{\binom{a}{i}\binom{n-a}{b-i}}{\binom{n}{b}}. \tag{5.1}$$

To convert the p-values into a distance we take the log transform with small added noise,

$$d_{AB} = \log_{10}(P_{AB} + 10^{-10}) + 10. \tag{5.2}$$

This yields a distance matrix D with distances scaled between 0 and 10. While this space does not explicitly reflect evolutionary divergence at a molecular level, it may be realistic at the protein level at which more complex types of genome evolution will have occurred.

We now compute the persistent homology of D. The barcode diagram is shown in Figure 5.3. The H_0 information represents hierarchical clustering and can be identically represented as a dendrogram. We show the dendrogram, restricting only to strains of order Caudovirales, in Figure 5.4. The strains are labeled by their taxonomic family: red for Myoviridae, blue for Siphoviridae, and green for Podoviridae. We can immediately see that the assigned taxonomic families are not consistent with the clustering based on protein information. However, there does appear to be some structure in which the taxonomic label is consistent within clusters of strains. Returning to the barcode diagram, we see substantial nontrivial homology in H_1 across all scales. This confirms the presence of mosaic exchange expected in phage genomes.

Focusing on order Caudovirales, for which the most data was present. We separately computed persistent homology for each of the three families. The barcode diagrams are shown in Figure 5.5. Computing the TOP score for each family, we have Myoviridae = 0.58, Siphoviridae = 1.14, and Podoviridae = 0.56.

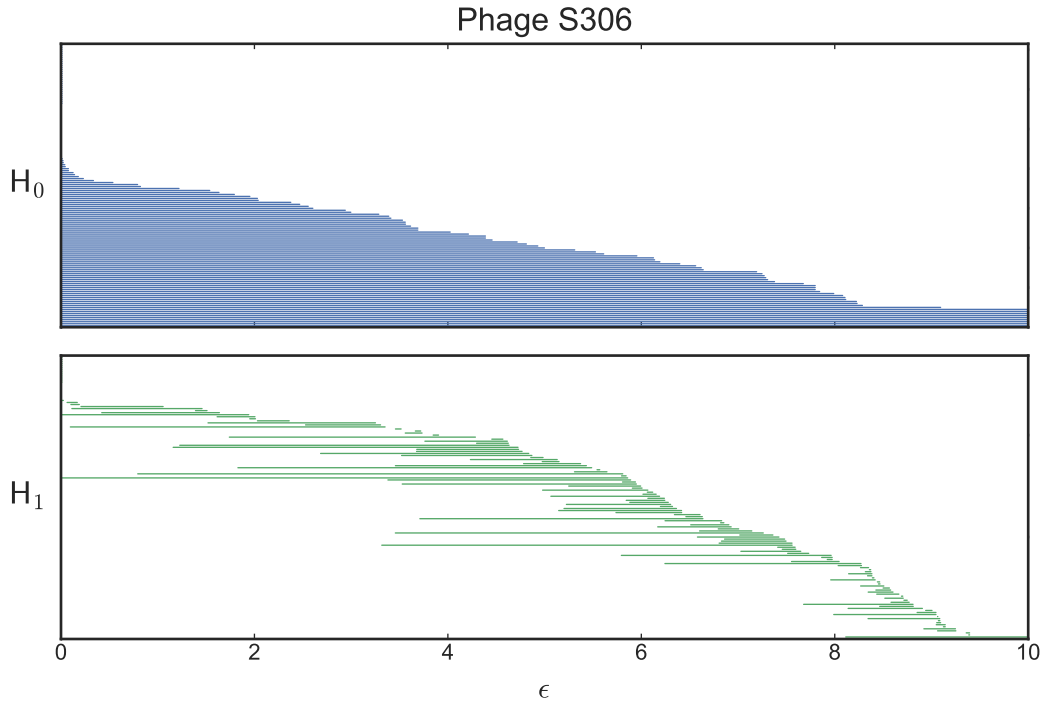


Figure 5.3: Bacteriophage Barcode Diagram using the S306 dataset

5.4 Representing Phage Relationships with Mapper

We used Ayasdi Mapper to construct a network representation of the phage phyletic profiles. The network was constructed using a Hamming metric on the phyletic matrix and a 2D filter function. The first filter was Metric PCA coordinate 1 with a resolution of 20 and a gain of 3.[3] The second filter was Metric PCA coordinate 2 with a resolution of 20 and a gain of 3. The equalize setting was used for both filter functions, which ensures that in the filtered space each bin has approximately the same number of points. This resulted in a network consisting of 201 nodes from the original 306 rows. The basic structure of the network is shown in Figure 5.6, where node color corresponds to the number of phages contained in the node. The network consists of one large connected component, two smaller connected components, and 21 singly connected nodes. The large connected component has local

[3]The parameter settings are in arbitrary units and tuned by hand to produce the most visually useful graph.

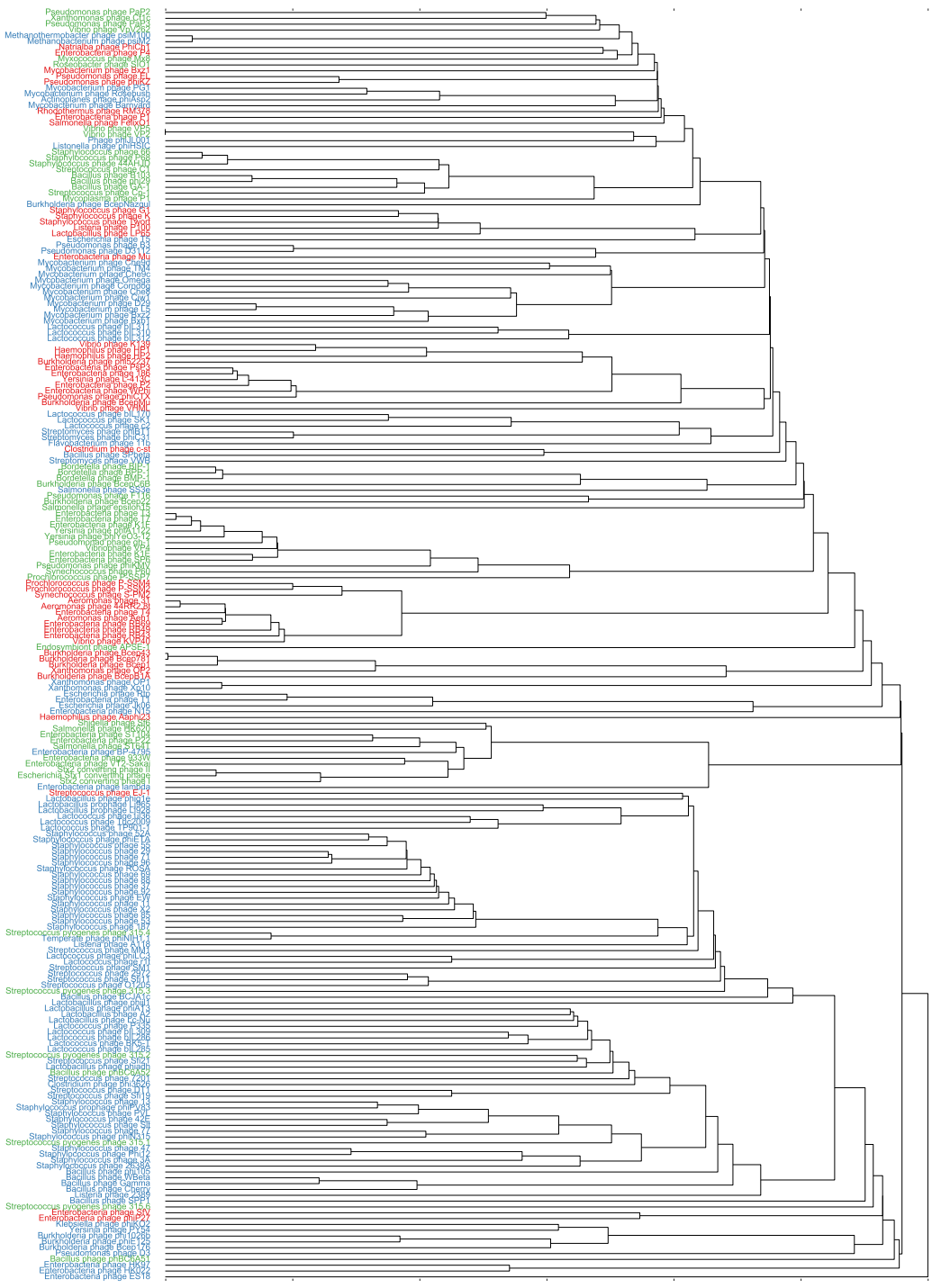


Figure 5.4: The dendrogram constructed from the H_0, restricted to bacteriophages of order Caudovirales. Myoviridae in red, Siphoviridae in blue, and Podoviridae in green. The family classifications are inconsistent with the hierarchical clustering.

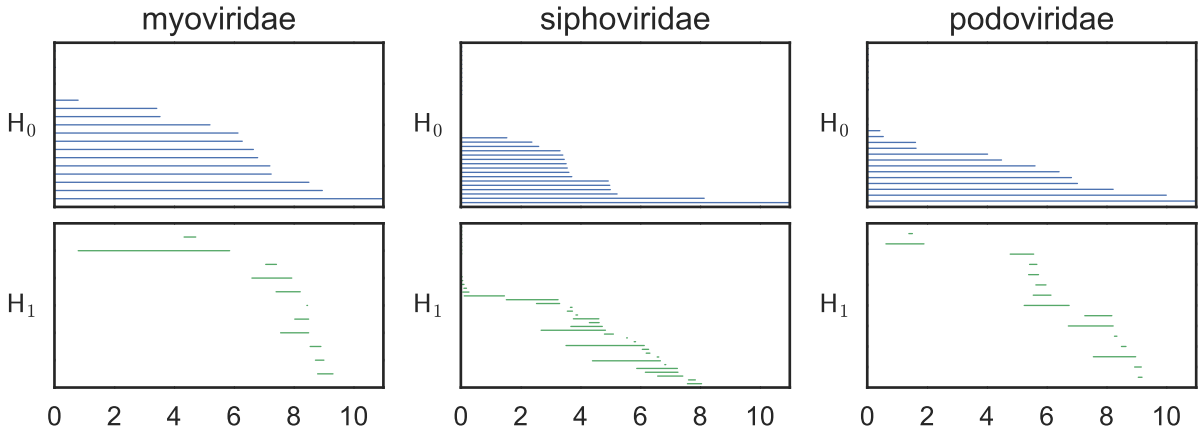


Figure 5.5: Barcode Diagrams for Families of Order Caudovirales, including Siphoviridae, Myoviridae, and Podoviridae.

regions of clustering, which will be considered further later.

We first examined how well the existing taxonomic classifications localized in the Mapper representation. If the taxonomy accurately reflects the molecular characteristics that were used to construct the network, we would expect to see strains belonging to the same level of hierarchy as localized together in the network, with minimal mixing between strains of different classification. We show the representation of the three families of order Caudovirales in Figure 5.7. Each node is colored by the proportion of rows from that family contained in the node.[4] We immediately see that each family is widely dispersed across the network. On closer examination, we see that the patterns of spread resemble those of the dendrogram in Figure 5.4, in that there are multiple clusters core clusters for each family. For example, the Myoviridae family has clusters in the bottom left and bottom right of the large component, and two singleton clusters. This roughly corresponds to the four clusters of Myoviridae in the H_0 dendrogram.

How strongly a particular classification is reflected by a network can be quantitatively measured using a modularity score [109]. Modularity was originally devised for identify-

[4]Recall that the nodes in a Mapper network can be composed of multiple nodes, depending on the parameters of the filter function used.

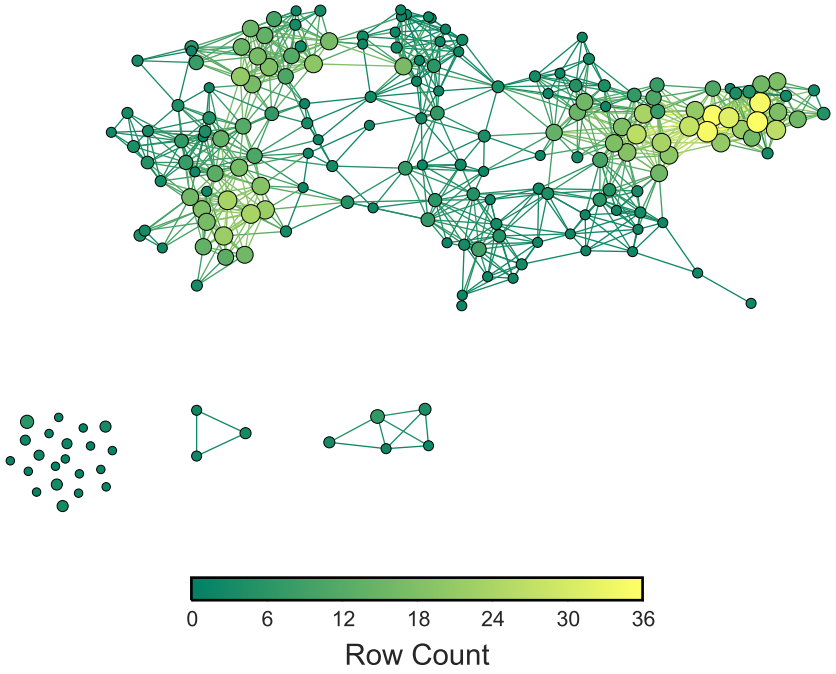


Figure 5.6: Phage Mapper Network. Network constructed using Mapper as implemented in Ayasdi Iris [7]. The network was constructed using a Hamming metric with a 2D Metric PCA filter function (resolution=20, gain=3, equalize). Nodes in the network represent clusters of phages and edges connect nodes that contain samples in common. Nodes are colored by the number of phages in each node.

ing community structure in networks. Intuitively, more tightly localized network divisions will have a higher modularity, while dispersed divisions will have a lower modularity. The standard definition for a two-class division is We use a modified form of modularity

$$Q = \frac{1}{m} \sum_{ij} A_{ij} s_i s_j \tag{5.3}$$

where m is the total number of edges in the network, A is the adjacency matrix of the network, and $s_i = \pm 1$ is the class membership of node i.[5] The modularity ranges between 0 and 1. We use a strict class membership, in which $s_i = 1$ for node i if any row in the now contains the annotation of interest. The modularity score for each family of Caudiovirales

[5]The standard definition of modularity includes a term measuring how tightly connected each module is.We are only interested in the localization of each modular and neglect this term.

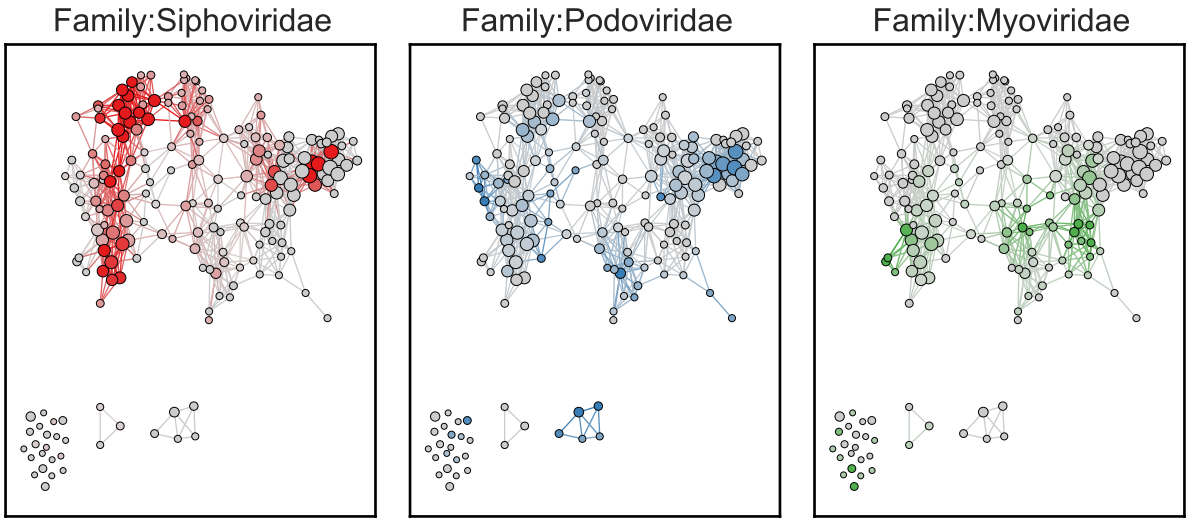


Figure 5.7: Taxonomic localization in the bacteriophage network. Network nodes are colored by presence of phages for each family in order Caudovirales.

is shown in Figure 5.8.

Second, we examined how well host correlated with network structure. We show this for the top six hosts represented in our dataset in Figure 5.9. While Enterobacteria has several pockets of representation within the network, phages are on average more strongly clustered by host than by taxonomy. This is consist with existing evidence that phages of similar host range have a common environment for reticulate exchange[89]. Modularity scores for the most dominant four hosts are shown in Figure 5.8. Staphylococcus has the highest defined modularity, which is consistent with the earlier reports about strong coupling and high levels of exchange between the Staphylococcus host and its viruses [41].

Finally, we clustered the network using the MCL graph clustering algorithm [55], as implemented in the Python MCLMarkovCluster package [87]. The MCL algorithm takes two input parameters which control the coarseness of the clustering: an expansion factor e and an inflation factor i. We set $e = 5$ and $i = 5$. Ignoring the singleton nodes, this resulted in eleven clusters, as shown in Figure 5.10. For each cluster, we used a hypergeometric test to identify particular protein families that were over- or under-represented in each cluster.

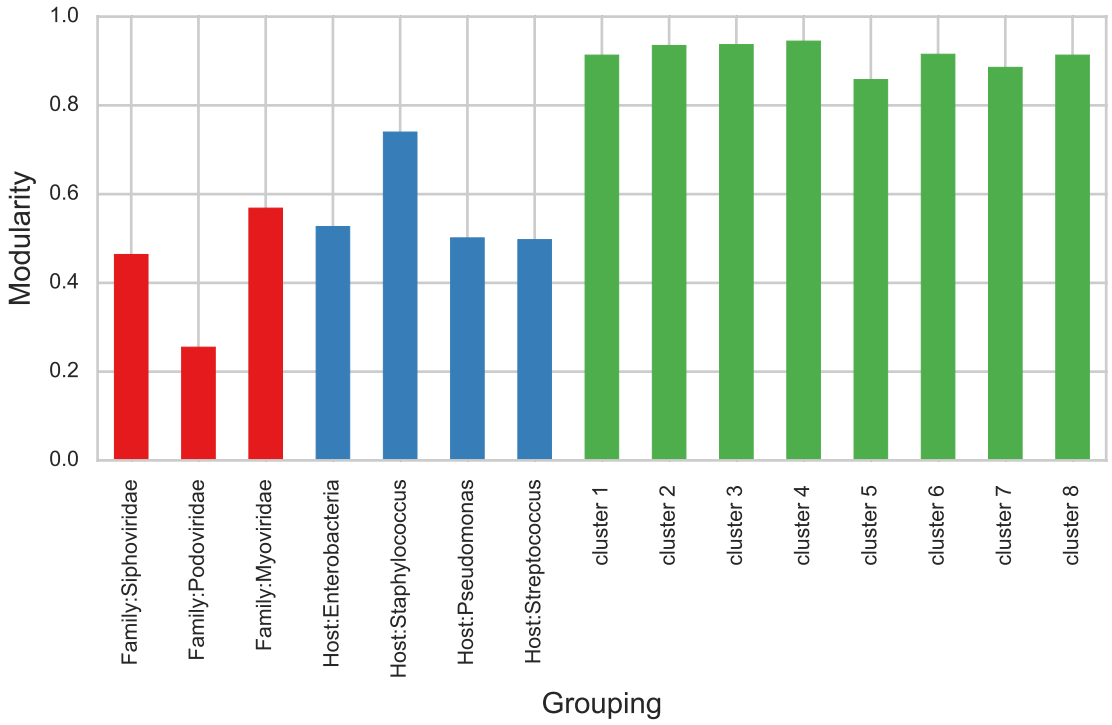


Figure 5.8: Modularity Scores for Different Divisions of the Phage Network. We show the modularities for divisions defined by taxonomic family and host range, as well as the clusters we identify using MCL.

After correcting for multiple testing, the protein families were most significantly associated with particular clusters are shown in Table 5.2.

5.5 Conclusions

In this chapter, we analyzed reticulate evolution in bacteriophages, using data from fully sequenced phage genomes represented as phyletic profiles measuring gene content. First, we used persistent homology to show that there are high levels of reticulate exchange across multiple taxonomic scales. Information in the H_0 barcode confirmed the inconsistency of the ICTV classification. Information in the H_1 barcode was used to compare levels of reticulate exchange among different phages. Second, we used Mapper to construct a network representation of phage molecular relationships. We examined how well different annotations,

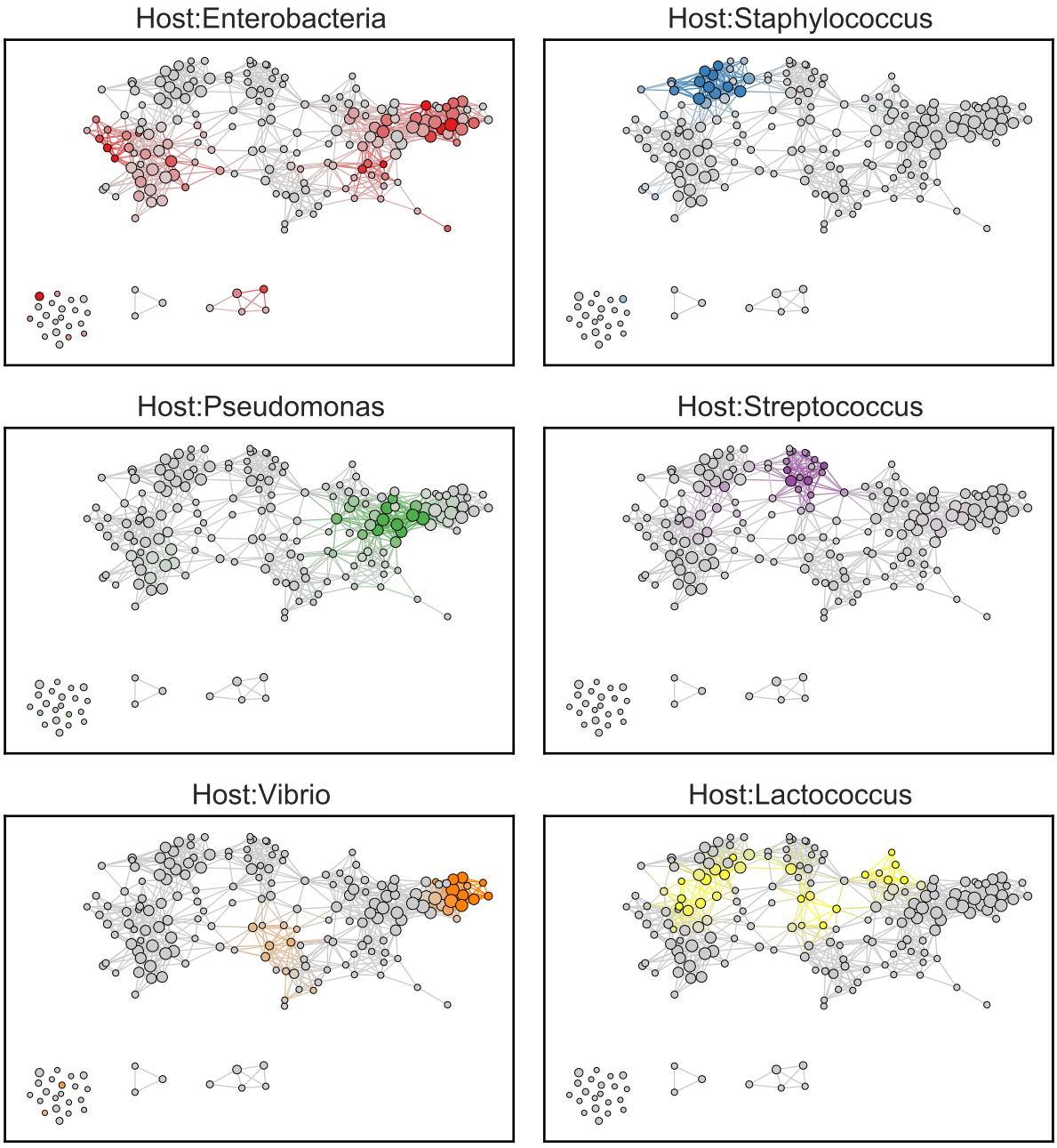


Figure 5.9: Host localization in the bacteriophage network. Compared to taxonomic family, phages are more tightly localized, reflecting the degree to which shared host range provides an environment for reticulate exchange.

Table 5.2: Phage Network MCL clustering annotations and representative protein families

Cluster	Protein Family	p-value	Function
Cluster 1	pf_0001	6.17e-25	tail tape measure protein
	pf_0002	6.81e-21	transcriptional repressor
	pf_0003	2.69e-16	tyrosine based integrase
	pf_0004	9.08e-12	DNA binding protein
	pf_0008	7.37e-10	unknown
	pf_0010	2.36e-08	terminase large subunit
	pf_0006	2.87e-08	NA
	pf_0164	5.57e-08	NA
	pf_0012	1.74e-07	DNA replication inititation protein
	pf_0015	2.42e-07	scaffolding protein
	pf_0187	2.70e-07	NA
	pf_0217	2.70e-07	NA
	pf_0013	3.35e-07	NA
	pf_0017	4.63e-07	portal protein
	pf_0007	8.84e-07	endolysin
Cluster 2	pf_0131	9.38e-13	NA
	pf_0279	2.33e-09	NA
	pf_0109	1.02e-08	NA
	pf_0434	4.38e-08	NA
	pf_0435	4.38e-08	NA
	pf_0436	4.38e-08	NA
	pf_0010	1.84e-07	terminase large subunit
	pf_0029	2.02e-07	major head protein
	pf_0009	2.54e-07	NA
	pf_0093	4.61e-07	NA
	pf_0512	5.47e-07	NA
	pf_0017	7.12e-07	portal protein
Cluster 3			
Cluster 4	pf_0049	3.44e-24	unknown
	pf_0043	3.44e-24	unknown
	pf_0053	3.86e-23	unknown
	pf_0052	3.86e-23	unknown
	pf_0007	6.19e-22	endolysin
	pf_0058	4.39e-21	unknown
	pf_0060	4.39e-21	unknown
	pf_0061	4.39e-21	unknown
	pf_0002	2.23e-20	transcriptional repressor
	pf_0067	4.46e-20	unknown
	pf_0063	4.46e-20	unknown
	pf_0012	1.12e-19	DNA replication inititation protein
	pf_0018	2.04e-19	tail protein
	pf_0072	4.38e-19	unknown
	pf_0004	1.82e-18	DNA binding protein
	pf_0020	5.91e-18	unknown
	pf_0087	3.86e-17	unknown
	pf_0042	5.24e-17	unknown
	pf_0001	1.54e-16	tail tape measure protein
	pf_0092	3.48e-16	unknown
Cluster 5	pf_0008	7.45e-07	unknown
Cluster 6	pf_0006	2.17e-18	NA
	pf_0057	8.74e-10	unknown
	pf_0142	3.84e-09	unknown
	pf_0082	1.79e-08	lysis protein
	pf_0165	2.09e-08	prohead
	pf_0188	1.12e-07	minor tail protein
	pf_0190	1.12e-07	tail
	pf_0189	1.12e-07	minor tail protein
	pf_0204	5.81e-07	unknown
Cluster 7	pf_0002	8.92e-11	transcriptional repressor
	pf_0121	2.46e-09	transcription factor
	pf_0016	3.36e-08	unknown
	pf_0122	6.52e-08	unknown
	pf_0156	3.35e-07	NA
	pf_0517	4.38e-07	NA
	pf_0004	5.26e-07	DNA binding protein
	pf_0135	5.97e-07	unknown
	pf_0176	9.43e-07	post-translational regulator
	pf_0178	9.43e-07	unknown
	pf_0177	9.43e-07	transcription anti-termination protein
	pf_0012	9.78e-07	DNA replication inititation protein
	pf_0324	1.06e-06	NA
	pf_0325	1.06e-06	NA
Cluster 8	pf_0497	1.24e-07	NA
	pf_0417	8.23e-07	NA
	pf_0002	9.70e-07	transcriptional repressor
Cluster 9	pf_0425	2.15e-14	NA
	pf_0424	2.15e-14	NA
	pf_0423	2.15e-14	NA
	pf_0422	2.15e-14	NA
	pf_0421	2.15e-14	NA
	pf_0420	2.15e-14	NA
	pf_0146	2.15e-14	NA
	pf_0366	1.72e-13	NA
	pf_0316	7.74e-13	NA
	pf_0315	7.74e-13	NA
	pf_0273	2.58e-12	NA
	pf_0274	2.58e-12	internal virion protein
	pf_0138	2.58e-12	head protein
	pf_0504	6.45e-12	NA
	pf_0503	6.45e-12	NA
	pf_0183	7.09e-12	NA
	pf_0184	1.70e-11	portal protein
	pf_0185	1.70e-11	tail protein
	pf_0163	1.70e-11	tail protein
	pf_0426	4.50e-11	NA

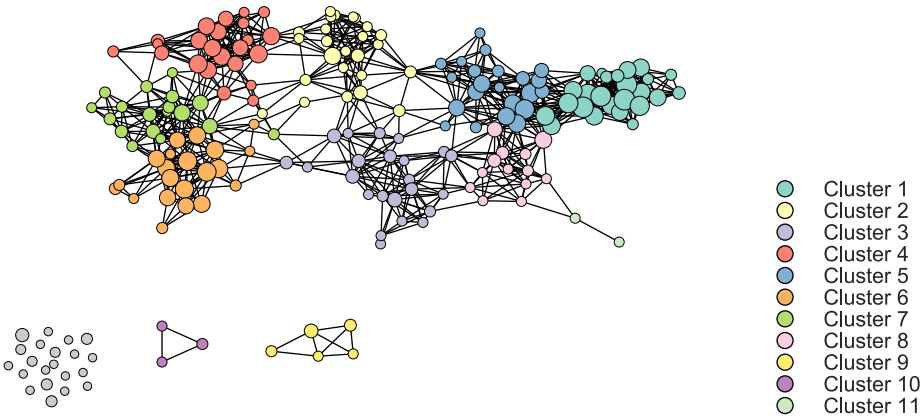


Figure 5.10: Phage Network with MCL Clustering. 11 nontrivial clusters are identified. In Table 5.2 we associate clusters with representative protein families.

including taxonomic classification and host range, localized on this network. We used a network clustering algorithm to identify communities of phages related by shared protein content, and identified protein families representative of each cluster. These clusters, while not explicitly reflecting potential phylogenetic trajectories, are more reflective of molecular similarity than existing morphological taxonomies, and can be used as a starting point for developing a more comprehensive picture of bacteriophage evolutionary dynamics. Further sequencing data will allows us to refine these clusters and provide a higher resolution

Chapter 6

Reassortment in Influenza Evolution

6.1 Introduction

In this chapter, we study influenza virus, a common human pathogen with a substantial burden on human health. Seasonal influenza epidemics have an annual mortality of between 250,000 and 500,000 [147]. Influenza pandemics, which have historically occurred roughly once every thirty years, can infect between 20-40% of the global population. For example, the Spanish influenza pandemic of 1918-1919 is estimated to have infected approximately 500 million people and lead to the death of between 50-100 million people [129]. This amounts to an infection of approximately 33% of the population and a case fatality ratio of 5-6% of global population.

The natural host reservoir of influenza is waterfowl. Within this reservoir, several distinct subtypes circulate. Subtypes are labeled by the antigenic type of two surface proteins, hemagglutinin (HA) and neuraminidase (NA).[1] There are presently eighteen types of HA (H1 to H18) and eleven types of NA (N1 to N11). Zoonotic adaptations have led to multiple introductions to human populations, which have resulted in both isolated outbreaks and

[1] An antigen is a molecule that elicits a host immune response. The adaptive immune system learns to recognize an antigenic type. Antigenic variation occurs when sufficient genomic alteration has occurred for a protein to evade the immune response.

sustained transmission [106].[2]

The evolution of influenza is punctuated by frequent reassortment. Reassortment occurs when two virus particles coinfect the same host cell, and is a consequence of influenza having a segmented genome. The result is viral progeny that carries genomic information from two independent parental strains. This mode of evolution is known as *antigenic shift*, because it can rapidly lead to antigenically distinct viral strains.[3] Antigenic shifts have historically been the cause of major pandemics, which can occur when novel surface proteins reassort with internal segments already adapted to the human host. Reassortments of this type led to Asian flu pandemic of 1957 and the Hong Kong flu pandemic of 1968 [94]. The 2009 H1N1 pandemic strain emerged from a triple reassortment between avian, swine, and human circulating strains [70, 124]. The pandemic had a global infection rate of between 11%-21% but a lower mortality rate than initially expected.[4] The 2013 H7N9 outbreak was caused by a triple reassortment of three distinct avian strains [33]. Traditionally, reassortments have been identified by hand by comparing phylogenetic trees constructed from different genomic segments [107].

Recent years have seen increased concerns about the pandemic potential for zoonotic adaptation of highly pathogenic strains of influenza. Of particular concern is H5N1, which has an estimated case fatality rate of 50% (449 deaths from 846 confirmed human cases) [148], but has so far not exhibited sustained person-to-person transmission [147]. These concerns underscore the need to efficiently characterize and represent reticulate evolution in influenza. Since the 2009 H1N1 pandemic, substantial effort has been put into collected fully sequenced influenza genomes. The NCBI Influenza Virus Resource now contains over 400,000 unique viral isolates [10]. The large quantity of genomic data that has been collected

[2]Understanding the genetic basis for host adaptation is an important and controversial research area. Our work in this area in collaboration with Yoshihiro Kawaoka is forthcoming [137].

[3]As opposed to *antigenic drift*, due to random mutation and genetic drift.

[4]The 2009 H1N1 pandemic is an excellent example of the delicate balance between virulence and transmissibility.

provides an ideal environment for studying reticulate evolution with high resolution.

6.2 Influenza Virology

Influenza is an enveloped single-stranded negative-sense RNA virus of family Orthomyxoviridae. The virus has a segmented genome with eight segments coding for eleven proteins. The genome length is approximately 13.5 kb. The viral structure is shown in Figure 6.1. The segments are typically ordered from longest to shortest and are detailed in Table 6.1. Of these segments, hemagglutinin (HA) and neuraminidase (NA) are the two most important. HA and NA form the two surface protein markers and On the surface of the viral particle are HA and NA. HA regulates host cell binding and entry into host epithelial cells. HA is the strongest determinant of host specificity: different hosts express different sialic acid types. Avian influenza binds to type 2-3 sialic acid receptors, while human influenza binds to type 2-6 sialic acid receptors. NA is the surface protein that cleaves the newly replicated virus particles from the cell surface. Together, HA and NA determine the strain subtype and are a primary marker of host transmission and specificity. Both are antigens. PA, PB1, and PB2 form a polymerase complex and are involved in viral replication. Mutations in these proteins can be among the most important in determining host adaptation and virulence. The remaining proteins, including NP, M1, M2, and NS1 are largely structural proteins involved in capsid formation and viral packaging.

6.3 Influenza Reassortment

We characterized reassortment in avian influenza using persistent homology. We first compiled an aligned dataset of 3,105 complete avian influenza genomes from the NIH Influenza Sequence Database. These sequences span in time from 1956 to 2012. We collected samples from all influenza subtypes. The distribution of collected HA types is shown in Figure XX.

Table 6.1: Influenza Genome Segments

Segment	Length (aa)	Name	Abbreviation	Proteins
1	xx	Polymerase basic 2	PB2	PB2
2	xx	Polymerase basic 1	PB1	PB1,PB1-F2
3	xx	Polymerase acidic	PA	PA
4	xx	Hemagglutinin	HA	HA
5	xx	Nucleoprotein	NP	NP
6	xx	Neuraminidase	NA	NA
7	xx	Matrix	M	M1,M2
8	xx	Nonstructural	NS	NS1

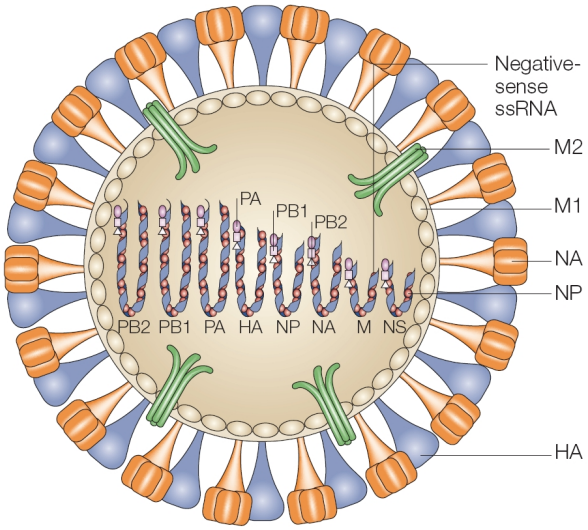


Figure 6.1: Structure of an influenza virus particle. Surface antigens HA and NA coat this surface and are involved in cell entry and exit. The surface capsid is formed from matrix proteins M1 and M2. PB1, PB2, and PA form a polymerase complex assisting in viral replication in the infected cell.

The majority of our sequences are of the H5 and H6 type, with a smaller proportion of H3, H7, and H9.

We first applied persistent homology to each genomic segment individually, as shown in Figure 6.2. Here we see broadly only zero-dimensional homology, consistent with no intra-segmental recombination. The presence of higher homology is likely due to back mutation, which is expected to be more common in viruses with high mutation rates and shorter genomes (i.e. the infinite sites model does not hold). However, an analysis of the concatenated full genome reveals a complex topology, with a large number of homological invariants in one and two dimensions.

In settings of vertical evolution, we can directly transform a filtration of 0-D simplicial complexes into an equivalent distance-based dendrogram. Fig. 2A represents the zero-dimensional topology of the hemagglutinin segment of avian influenza viruses. The zero-dimensional generators at higher genetic distances indicate the major clusters, coinciding with the antigenic subtypes H1-H16. From the bar sizes of the barcode plot, we can create a dendrogram that recapitulates classic phylogenetic analyses (Fig. 2B). Only when segments are concatenated does persistent homology indicate that reassortment precludes phylogenetic analysis (Fig. 2C). These results show that persistent homology can detect pervasive reassortment in influenza. Estimating ICR from one-dimensional homology provides a lower-bound on reassortment rate in influenza. We calculate an ICR of <1 event per year for classic H1N1 swine and H3N2 human influenza viruses, supported by previous phylogenetic estimates. In contrast, we calculate a high reassortment rate of 22.16 events per year for avian influenza A. This difference could be explained by the high diversity and frequent co-infection of avian viruses and correlates with the high proportion of potential avian reassortants reported in previous studies.

To illustrate how higher-dimensional topology captures reassortments, we analyzed 1,000 human H3N2 genomes and identified three generators of one-dimensional homology when joining the PB2 and HA segments. As an example, the [G3] generator with the longest

bar (Dataset S1, Table S5) is represented by an oriented one-dimensional irreducible cycle, implying at least one reassortment involving PB2 and HA of the isolates or their ancestors. The number of sequences in the generator serves as an upper bound on the number of candidate reassortants. Simple observation of the resulting sequence alignment reveals two divergent allelic patterns between informative sites in PB2 and HA, as reflected in incongruent trees (SI Appendix, Fig. S8 A and B) and reticulate cycles of the phylogenetic network (SI Appendix, Fig. S8C).

One-dimensional ICR provides a lower-bound estimate of reassortment rate (SI Appendix, Fig. S10B). We calculate ICR < 1 event per year for classic H1N1 swine and H3N2 human influenza, supported by previous phylogenetic estimates [97, 71]. In contrast, we calculate a high rate of 22.16 reassortments per year for avian influenza A (Dataset S1, Table S16). This difference could be explained by the high diversity and frequent coinfection of avian viruses [95] and correlates with the high proportion of avian reassortants reported in previous studies [50].

We used mapper to visualize the relationships in our influenza dataset. A series of mapper networks is shown in Figure 6.4. The networks were generated using a Hamming metric and the first and second MDS component as a 2d lens. In each subfigure we color the network by the prevalence of each HA subtype in the particular node. It is interesting to note how this is different than a traditional phylogeny based on HA.

6.4 Nonrandom Association of Genome Segments

We observed nonrandom association of flu segments. Statistical inference on the loops corresponding to reassortments identified segments that tend to co-segregate with each other during reassortment.In particular, polymerases co-segregate, while genes coding for envelope and capsid proteins show independent reassortment patterns. Cosegregation of polymerases suggests that effective proteinprotein interaction between the polymerase complex and the

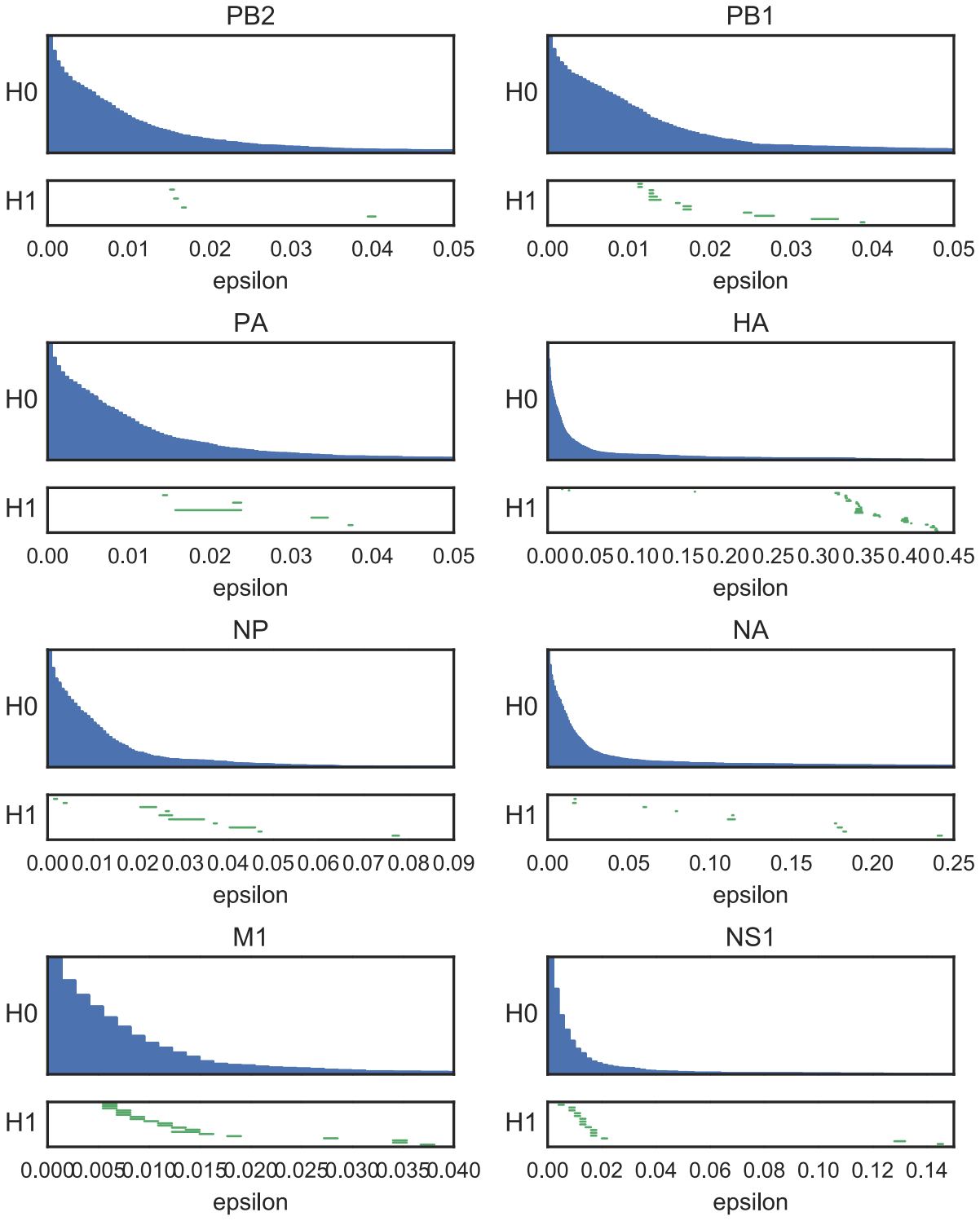


Figure 6.2: Influenza Segment Barcodes. Very little H_1 present.

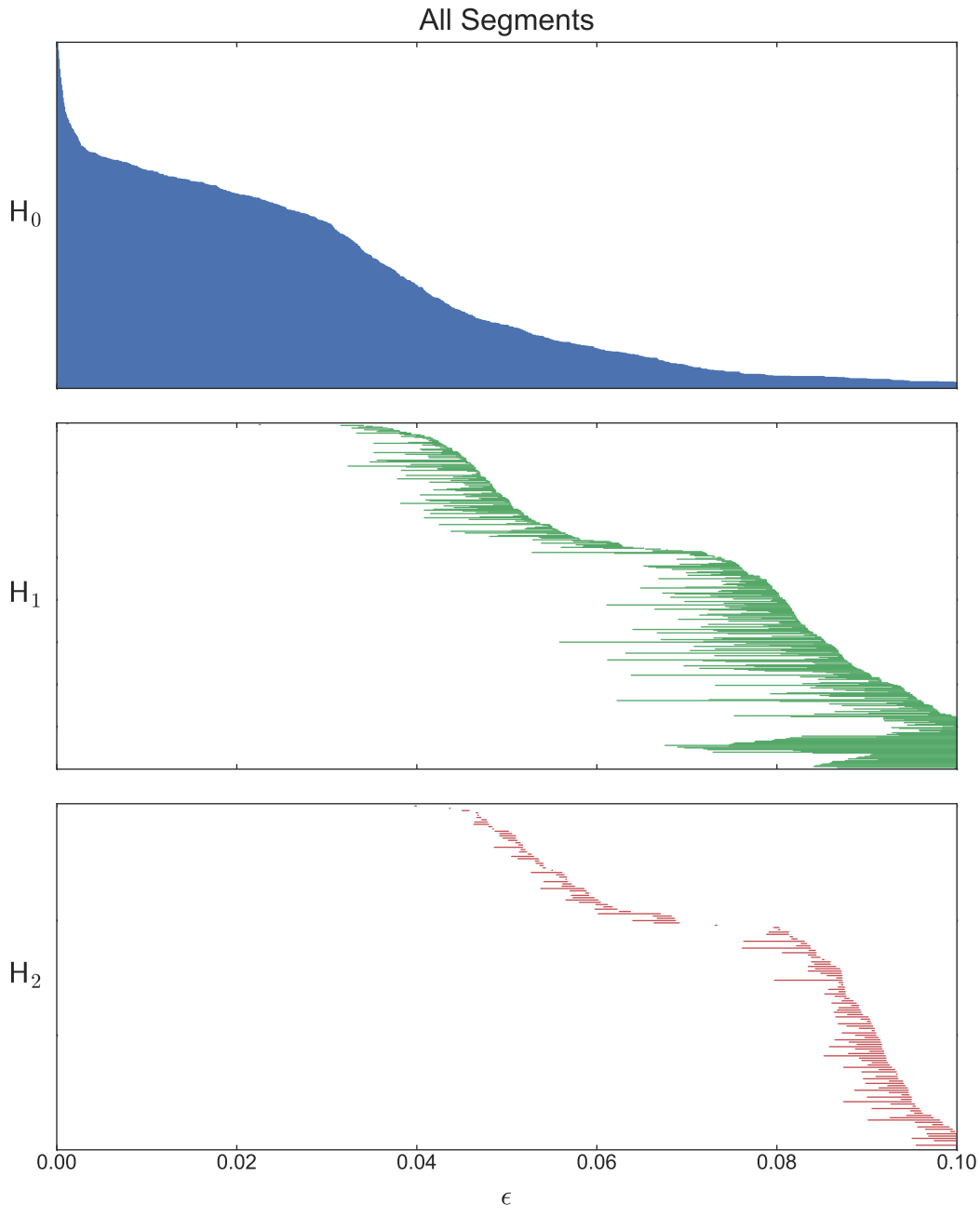


Figure 6.3: Influenza Concatenated Genome Barcode

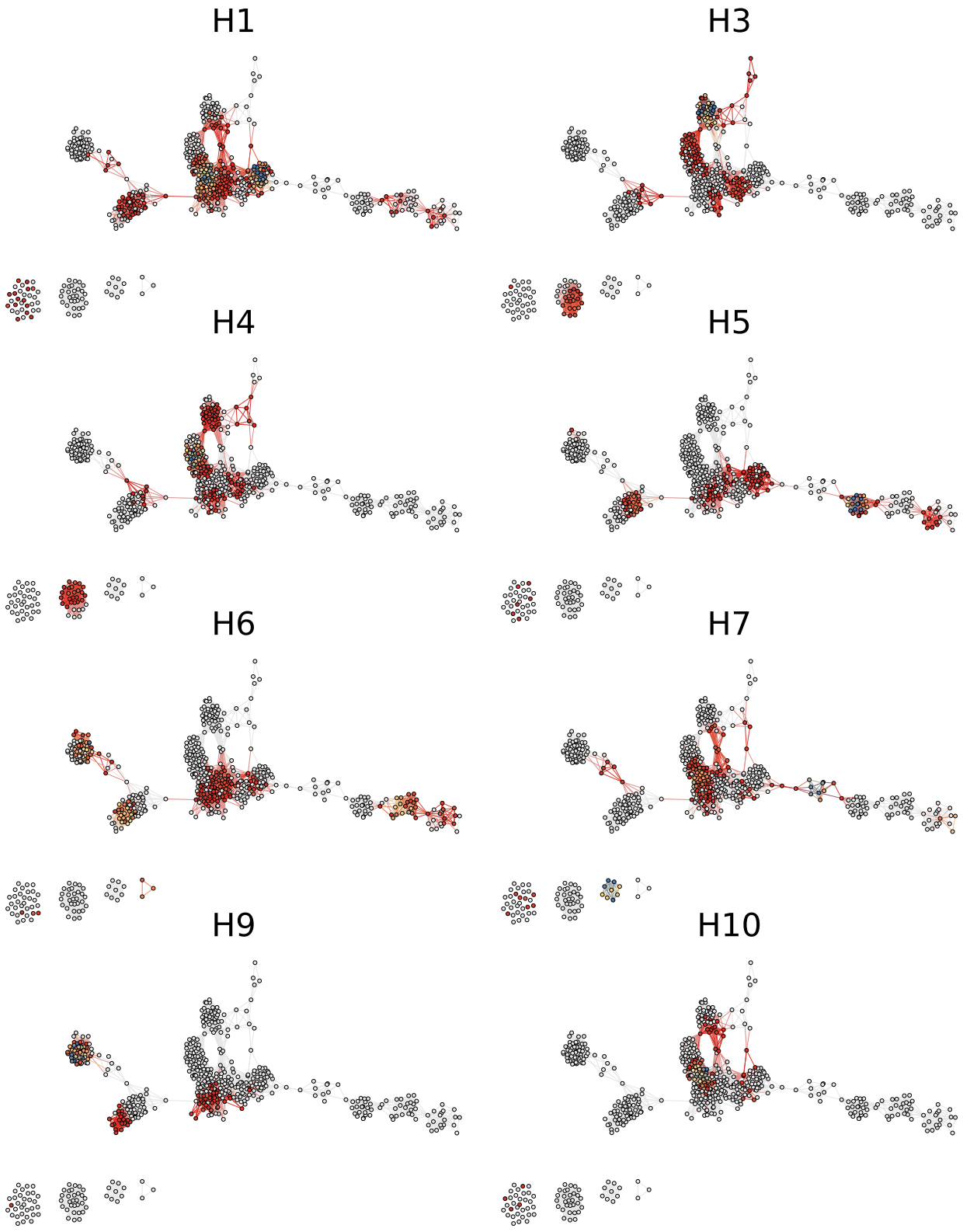


Figure 6.4: Influenza Networks By HA Subtype. The networks were generated using Ayasdi.

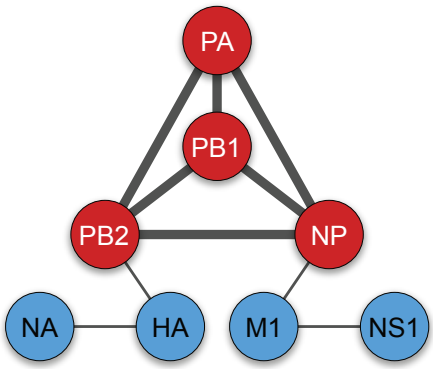


Figure 6.5: Influenza Nonrandom Reassortment

NP protein constrain reassortment.

Although previous phylogenetic studies confirmed a high reassortment rate in avian influenza, none has identified a clear pattern of gene segment association [50]. To determine whether any segments cosegregate more than expected by chance, we applied persistent homology to avian influenza. We first considered all pairs of concatenated segments and estimated the number of reassortments by $b1$. We then ascertained the significance of observing a number of reassortments between each pair of segments given the total estimate of reassortments in the concatenated genome (SI Appendix, Supplementary Text). Analysis of avian influenza reveals a statistically significant configuration of four cosegregating segments: polymerase basic 2 (PB2), polymerase basic 1 (PB1), polymerase acidic (PA), and nucleoprotein (NP) (Fig. 3D). Interestingly, this pattern mimics previous in vitro results that suggest that effective protein-protein interaction between the polymerase complex and the NP protein constrain reassortment [95].

6.5 Multiscale Flu Reassortment

We computed persistent homology on an aligned dataset of 3,105 avian influenza sequences across the seven major HA subtypes. The persistence diagram is shown in Figure 6.6, along

with density estimates for the birth and death distributions. Both birth and death times appear strongly bimodal, unlike in the coalescent simulations, which were strictly unimodal. This suggests two distinct scales of topological structure. Using the representative cycles output by Dionysus on a subset of this data, we classified features as intrasubtype (involving one HA subtype) and intersubtype (involving multiple HA subtypes). The H_1 barcode diagram for this data is shown in the Figure 6.6 inset. Intrasubtype features, in blue, occur at an earlier filtration scale than intersubtype features, in green. The multiscale topological approach of persistent homology can distinguish biological events occuring at different genetic scales.

We isolated the two peaks and estimated two recombination rates: an intrasubtype $\rho_1 = 9.68$, and an intersubtype $\rho_2 = 21.43$. We conclude that intersubtype recombination occurs at a rate over twice that of intrasubtype recombination, however a genetic barrier exists that maintains distinct subtype populations. The nature of this barrier warrants further study. This illustrates a real world example in which multiscale topological structure can be captured by persistent homology and given biological interpretation.

6.6 Prediction of Host Specific Residues

In this section, we describe work in prediction of host specific residues using machine learning approaches. Host specific residues are important for viral surveillance in order to predict possible outbreaks. We describe here two methods and include preliminary validation from our collaborator in Wisconsin.

6.7 Conclusions

The segmented nature of the influenza genome makes it an ideal Coinfection of a single cell by Coinfection of a single cell with multiple strains of influenza can lead to genomic reassortment. Reassortments can lead to novel pandemics. Therefore it is important that

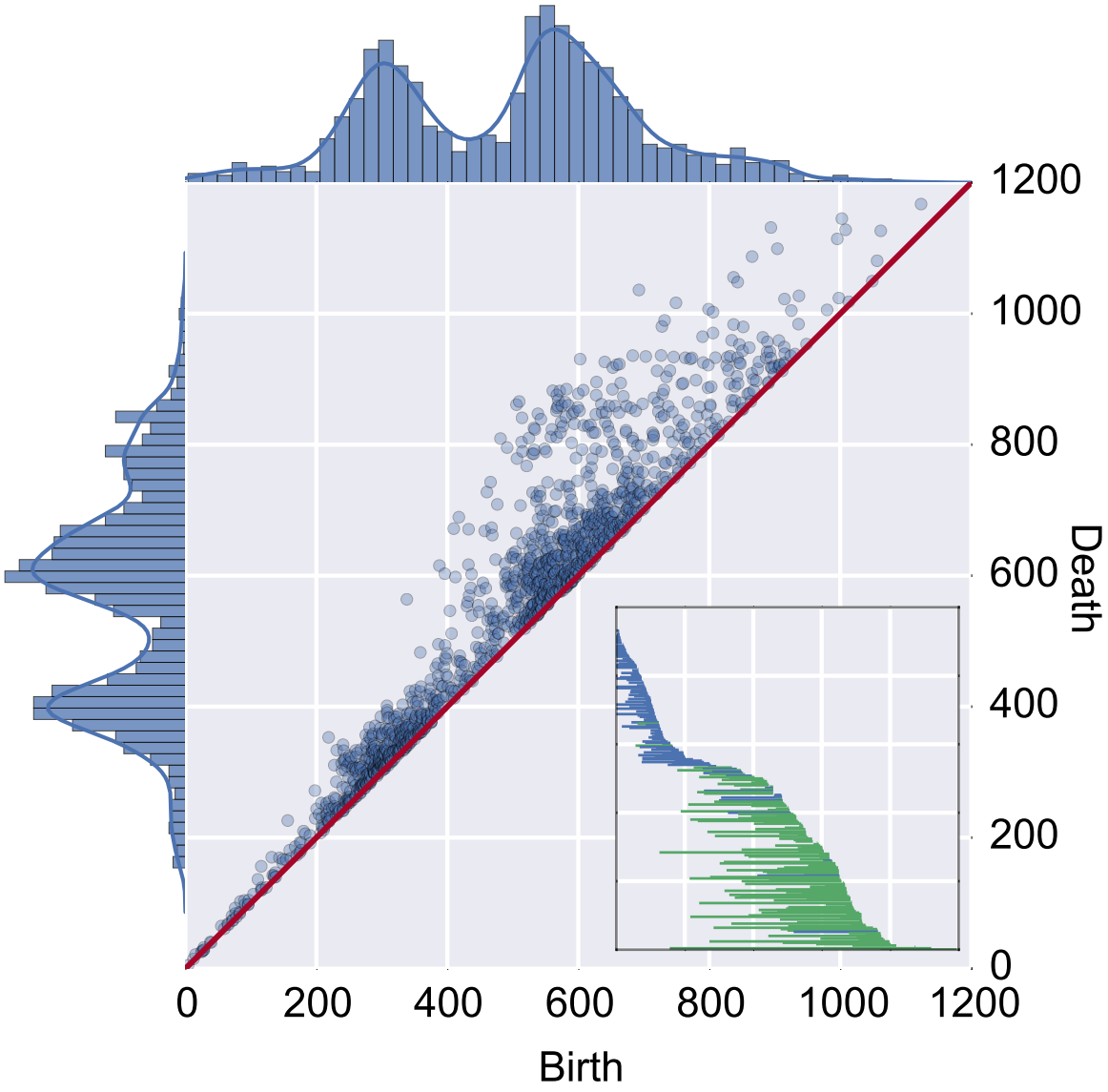


Figure 6.6: The H_1 persistence diagram computed from an avian influenza dataset. On the top and left are plotted the marginal distributions of birth and death times, along with a density estimate for each distribution. The bimodality indicates two scales of topological structure. Inset: The barcode diagram for a subset of this data. Blue bars have representative cycles involving only one subtype, green bars have cycles involving multiple subtypes.

methods to characterize reassortment be developed. In this chapter we have applied methods from TDA to characterize reassortment in influenza. Using our approach, we have confirmed that intrasegmental recombination does not occur. We have estimated reassortment rates. We have estimated recombination rates. Further, from the persistence diagram we identified a bimodal presence of H_1 invariants. This suggests a genetic barrier maintaining subtype diversity.

Chapter 7

Reticulate Evolution in Pathogenic Bacteria

7.1 Introduction

Pathogenic bacteria can lead to severe infection and mortality and present an enormous burden on human populations and public health systems. One of the achievements of twentieth century medicine was the development of a wide range of antibiotic drugs to control and contain the spread of pathogenic bacteria, leading to vastly increased life expectancies and global economic development. However, rapidly rising levels of multidrug antibiotic resistance in several common pathogens, including *Escherichia coli*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, and *Neisseria gonorrhoea*, is recognized as a pressing global issue with near-term consequences [108, 133, 146]. The threat of a post-antibiotic 21st century is serious, and new methods to characterize and monitor the spread of resistance are urgently needed.

Antibiotic resistance can be acquired through point mutation or through horizontal transfer of resistance genes. Horizontal exchange occurs when a donor bacteria transmits foreign DNA into a genetically distinct bacteria strain. As discussed in Chapter 2, three mechanisms

of horizontal transfer have been identified, depending on the route by which foreign DNA is acquired [111]. Foreign DNA can be acquired via uptake from an external environment (transformation), via viral-mediated processes (transduction), or via direct cell-to-cell contact between bacterial strains (conjugation). Resistance genes can be transferred between strains of the same species, or can be acquired from different species in the same environment. While the former is generally more common, an example of the latter is the phage-mediated acquisition of Shiga toxin in *E. coli* in Germany in 2011 [118]. Elements of the bacterial genome that show evidence of foreign origin are called genomic islands, and are of particular concern when associated with phenotypic effects such as virulence or antibiotic resistance.

In this chapter we explore topics relating to horizontal gene transfer in bacteria and the emergence of antibiotic resistance in pathogenic strains. We show that TDA can not only quantify gene transfer events, but also characterize the scale of gene transfer. The scale of recombination can be measured from the distribution of birth times of the H_1 invariants in the barcode diagram. It has been shown that recombination rates decrease with increasing sequence divergence [61]. We characterize the rate and scale of intraspecies recombination in several pathogenic bacteria of public health concern. We select a set of pathogenic bacteria that are of public health interest based on a recently released World Health Organization (WHO) report on antimicrobial resistance [146]. Using persistent homology, we characterize the rate and scale of recombination in the core genome using multilocus sequence data. To extend our characterization to the whole genome, we use protein family annotations as a proxy for sequence composition. This allows us to compute a similarity matrix between strains. Comparing persistence diagrams gives us information about the relative scales of gene transfer at arbitrary loci. The species selected for study and the sample sizes in each analysis are specified in Table 7.1. Next, we explore the spread of antibiotic resistance genes in *S. aureus* using Mapper, an algorithm for partial clustering and visualization of high dimensional data [123]. We identify two major populations of *S. aureus*, and observe one cluster with strong enrichment for the antibiotic resistance gene *mecA*. Importantly,

Table 7.1: Pathogenic bacteria selected for study and sample sizes in each analysis.

Species	MLST profiles	PATRIC profiles
Campylobacter jejuni	7216	91
Escherichia coli	616	1621
Enterococcus faecalis	532	301
Haemphilus influenzae	1354	22
Helicobacter pylori	2759	366
Klebsiella pneumoniae	1579	161
Neisseria spp.	10802	234
Pseudomonas aeruginosa	1757	181
Staphylococcus aureus	2650	461
Salmonella enterica	1716	638
Streptococcus pneumoniae	9626	293
Streptococcus pyogenes	627	48

resistance appears to be increasingly spreading in the second population. Finally, we consider the risk of lateral transfer of resistance genes from the human microbiome into an antibiotic sensitive strain, using β-Lactam resistence as an example. In this environment, benign bacterial strains can harbor known resistance genes. We use a network analysis to visualize the spread of antibiotic resistance gene *mecA* into nonnative phyla. Each indidivual has a unique microbiome, and we speculate that microbiome typing of this sort may useful in developing personalized antibiotic therapies. These results suggest an important role for topological data mining of -omics scale data in clinical applications and personalized medicine.

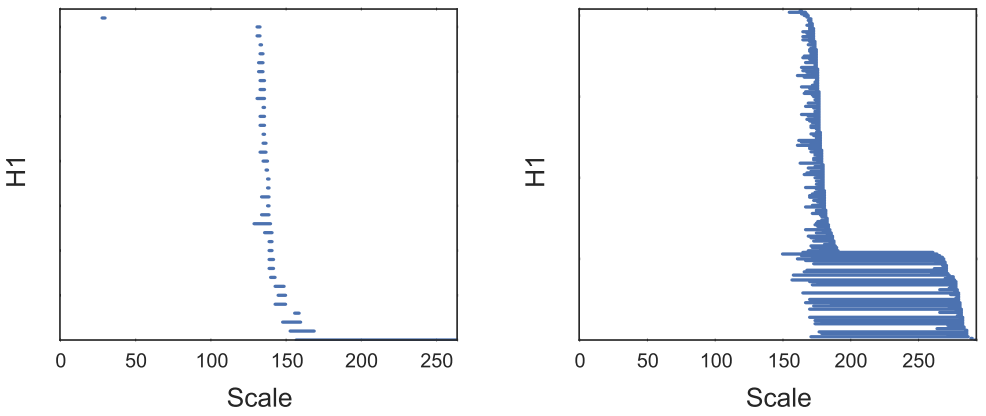
7.2 Evolutionary Scales of Recombination in the Core Genome

Multilocus sequence typing (MLST) data was used to examine scales of recombination in the core bacterial genome. MLST is a method of rapidly assigning a sequence profile to a sample bacterial strain. For each species, a predetermined set of loci on a small number of housekeeping genes are selected as representative of the core genome of the species. As

new strains are sequenced, they can be annotated with a profile corresponding to the type at each locus. If a sample has a previously unseen type at a given locus, it is appended to the list of types at that locus. Large online databases have curated MLST data from labs around the world; significant pathogens can have several thousand typed strains (over 10,000 in the case of *Neisseria spp.*). Because different species will be typed at different loci, examining direct interspecies genetic exchange with this data is unfeasible, however MLST provides a large quantity of data with which to examine intraspecies exchange in the core genome. However, because the selected loci are generally all housekeeping genes, this type of recombination analysis will tell you only about genetic exchane in the core genome. Mobile genetic elements may have a separate rates of exchange.

We investigate genetic exchange in the twelve pathogens using MLST data from PubMLST [79]. For each strain, a pseudogenome can be constructed by concatenating the typed sequence at each locus. Using a Hamming metric, we construct a pairwise distance matrix between strains and compute persistent homology on the resulting metric space. Because of the large number of sample strains, we employ a Lazy Witness complex with 250 landmark points and $\nu = 0$ [40]. The computation is performed using javaplex [130]. An example of our output is shown in Figure 7.1, where we plot the H_1 barcode diagrams for *K. pneumoniae* and *S. enterica*. The two species have distinct recombination profiles, characterized by the range of recombinations: *K. pneumoniae* recombines at only one short-lived scale, while *S. enterica* recombines both at the short-lived scale and a longer-lived scale. We repeat this analysis for each species, and plot the results as a persistence diagram in Figure 7.2. Among the bulk of pathogens there appears to be three major scales of recombination, a short-lived scale at intermediate distances, a longer-lived scale at intermediate distances, and a short-lived scale at longer distances. *H. polyori* is a clear outlier, tending to recombine at scales significantly lower than the other pathogens.

We define a relative rate of recombination by counting the number of H_1 loops across the filtration and dividing by the number of samples for that species. The results are shown in



(a) *Klebsiella pneumoniae* (b) *Salmonella enterica*

Figure 7.1: Barcode diagrams reflect different scales of core genomic exchange in *K. pneumoniae* and *S. enterica*.

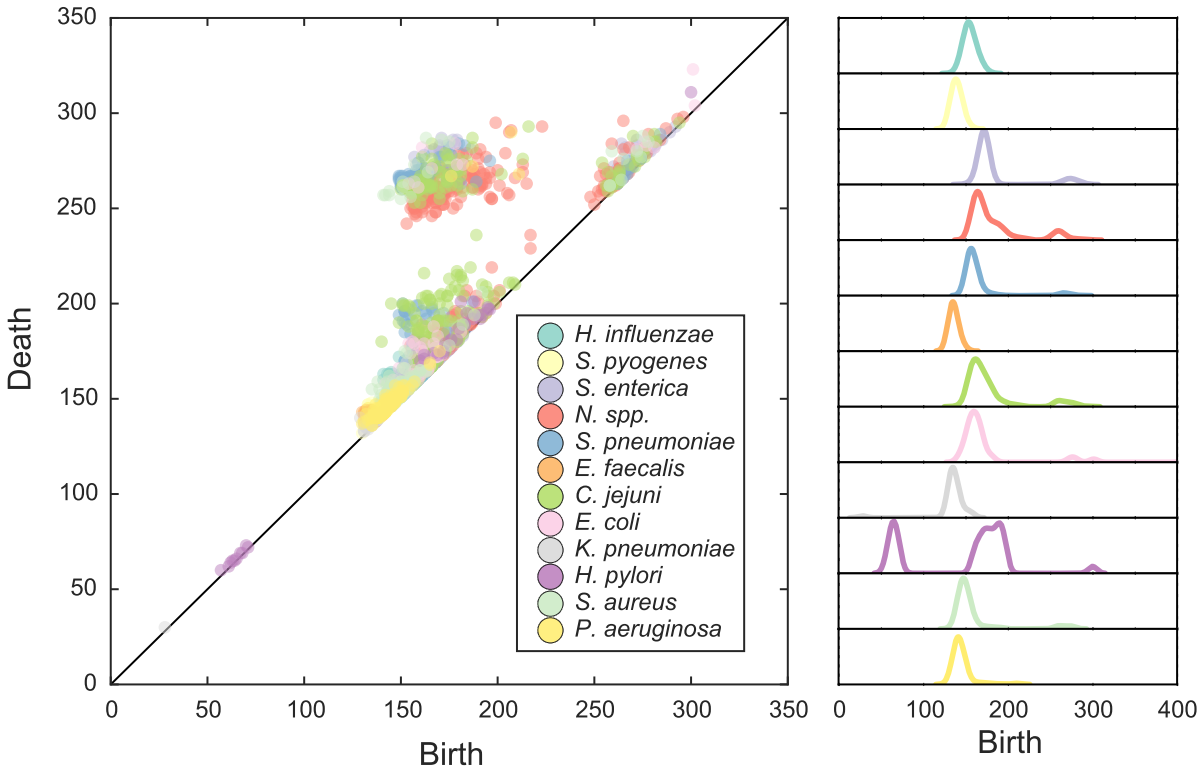


Figure 7.2: The H_1 persistence diagram for the twelve pathogenic strains selected for this study using MLST profile data. There are three broad scales of recombination. To the right is the birth time distribution for each strain. *H. pylori* has an earlier scale of recombination not present in the other species.

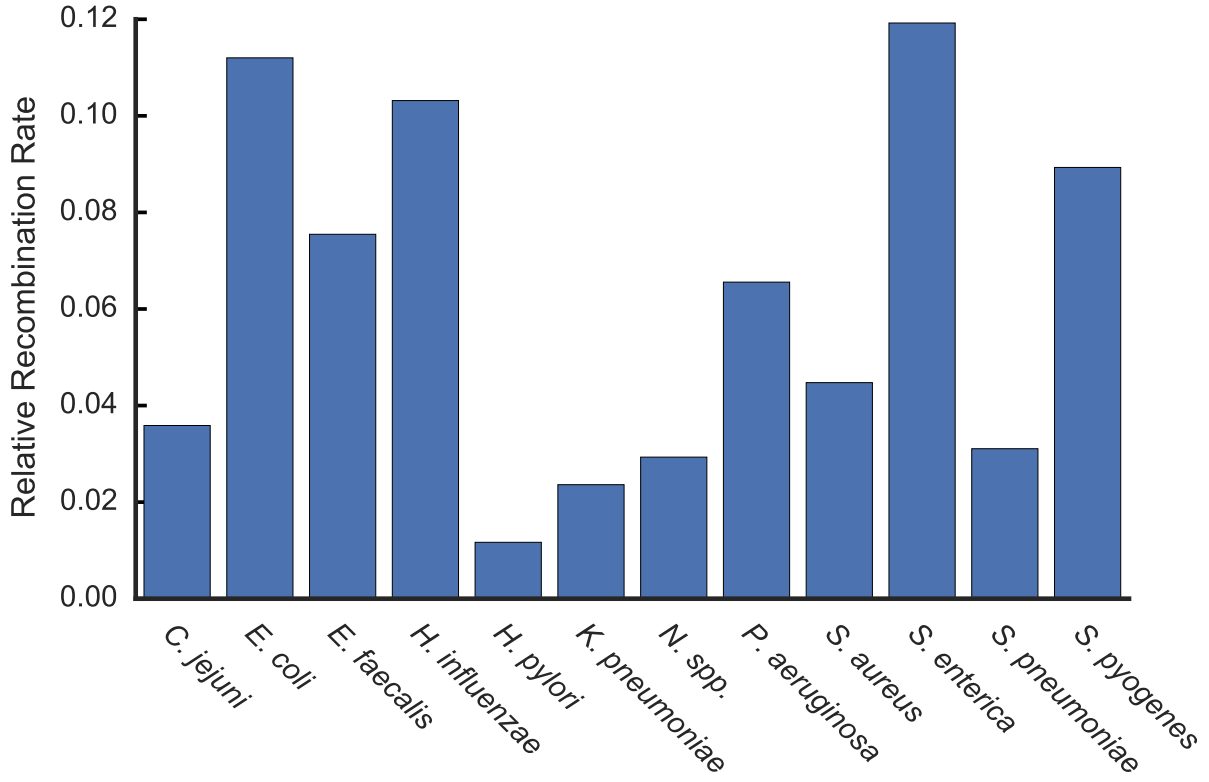


Figure 7.3: Relative recombination rates computed by persistent homology from MLST profile data.

Figure 7.3, where we observe that different species can have vastly different recombination profiles. For example, *S. enterica* and *E. coli* have the highest recombination rates, while *H. pylori* is substantially lower than the others. Coupled with the smaller scale of recombinations suggests that the *H. pylori* core genome is relatively resistant to recombination except within closely related strains.

7.3 Protein Families as a Proxy for Genome Wide Reticulation

Protein family annotations cluster proteins into sets of isofunctional homologs, i.e., clusters of proteins with both similar sequence composition and similar function. A particular strain

is represented as a binary vector indicating the presence or absence of a given protein family. Correlations between strains can reveal genome-wide patterns of genetic exchange, unlike the MLST data which can only provide evidence of exchange in the core genome. We use the FigFam protein annotations in the Pathosystems Resource Institute Center (PATRIC) database because of the breadth of pathogenic strain coverage and depth of genomic annotations [140]. The FigFam annotation scheme consists of over 100,000 protein families curated from over 950,000 unique proteins [101].

For each strain we compute a transformation into FigFam space. We transform into this space because the frequency of genome rearrangements and differences in mobile genetic elements makes whole genome alignments unreliable, even for strains within the same species. As justification for performing this step, it has been shown experimentally that recombination rates decrease with increasing genetic distance [61]. After transforming, we construct a strain-strain correlation matrix and compute the persistent homology in this space. In Figure 7.4 we show the persistence diagram relating the structure and scale between different species. We find that different species have a much more diverse topological structure in this space than in MLST space, and a wide variety of recombination scales. The large scales of exchange in *H. influenzae* suggest it can regularly acquire novel genetic material from distantly related strains.

7.4 Antibiotic Resistance in *Staphylococcus aureus*

S. aureus is a gram positive bacteria commonly found in the nostrils and upper respiratory tract. Certain strains can cause severe infection in high-risk populations, particulary in the hospital setting. The emergence of antibiotic resistant *S. aureus* is therefore of significant clinical concern. Methicillin resistant *S. aureus* (MRSA) strains are resistant to β-lactam antibiotics including penicillin and cephalosporin. Resistance is conferred by the gene *mecA*, an element of the Staphyloccoccal cassette chromosome mec (*SCCmec*). *mecA* codes for a

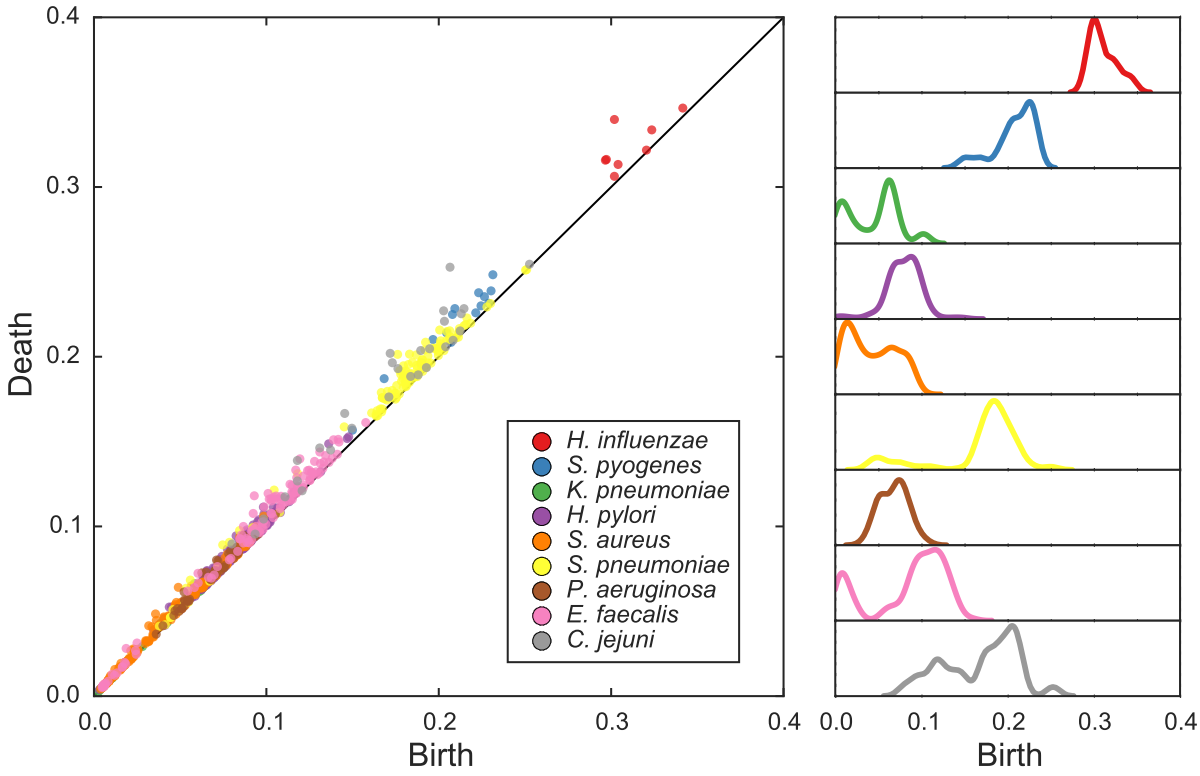


Figure 7.4: Persistence diagram for a subset of pathogenic bacteria, computed using the FigFam annotations compiled in PATRIC. Compared to the MLST persistence diagram, the Figfam diagram has a more diverse scale of topological structure.

dysfunctional penicillin-binding protein 2a (PBP2a), which inhibits β-lactam antibiotic binding, the primary mechanism of action [77]. Of substantial clinical importance are methods for characterizing the spread of MRSA within the *S. aureus* population.

To address this question, we use the FigFam annotations in PATRIC, as described in the previous section. PATRIC contains genomic annotations for 461 strains of *S. aureus*, collectively spanning 3,578 protein families. We perform a clustering analysis using the Mapper algorithm as implemented in Ayasdi Iris [7]. Principal and second metric singular value decomposition are used as filter functions, with a 4x gain and an equalized resolution of 30. This results in a graph structure with two large clusters, with a smaller bridge connecting the two, as shown in Figure 7.5. The two clusters are consistent with previous phylogenetic studies using multilocus sequence data to identify two major population groups [35].

Of the 461 *S. aureus* strains in PATRIC, 142 carry the *mecA* gene. When we color nodes in the network based on an enrichment for the presence of *mecA*, we observe a much stronger enrichment in one of the two clusters. This suggests that β-lactam resistance has already begun to dominate in that clade, likely due to selective pressures. More strikingly, we observe that while *mecA* enrichment is not as strong in the second cluster, there is a distinct path of enrichment emanating along the connecting bridge between the two clusters and into the less enriched cluster. This suggests the hypothesis that antibiotic resistance has spread from the first cluster into the second cluster via strains intermediate to the two, and will likely continue to be selected for in the second cluster.

7.5 Microbiome as a Reservoir of Antibiotic Resistance Genes

While antibiotic resistance can be acquired through gene exchange between strains of the same species, it is also possible for gene exchange to occur between distantly related species. It has been recognized that an individual's microbiome, the set of microorganisms that exist symbiotically within a human host, can act as a reservoir of antimicrobial resistance genes [125, 114]. It is of substantial clinical interest to characterize to what extent an individual's microbiome may pose a risk for a pathogenic bacteria acquiring a resistance gene through lateral transfer.

To address this question, we use data from the Human Microbiome Project (HMP), a major research initiative performing metagenomic characterization of hundreds of healthy human microbiomes [132]. The HMP has defined a set of reference strains that have been observed in human microbiomes. We collect FigFam annotations from PATRIC for the reference strain list in the gastrointestinal tract. We focus on the gastrointestinal tract because it is an isolated environment and likely to undergo higher rates of exchange than other anatomic regions. Of the 717 reference strains, 321 had FigFam annotations. We computed

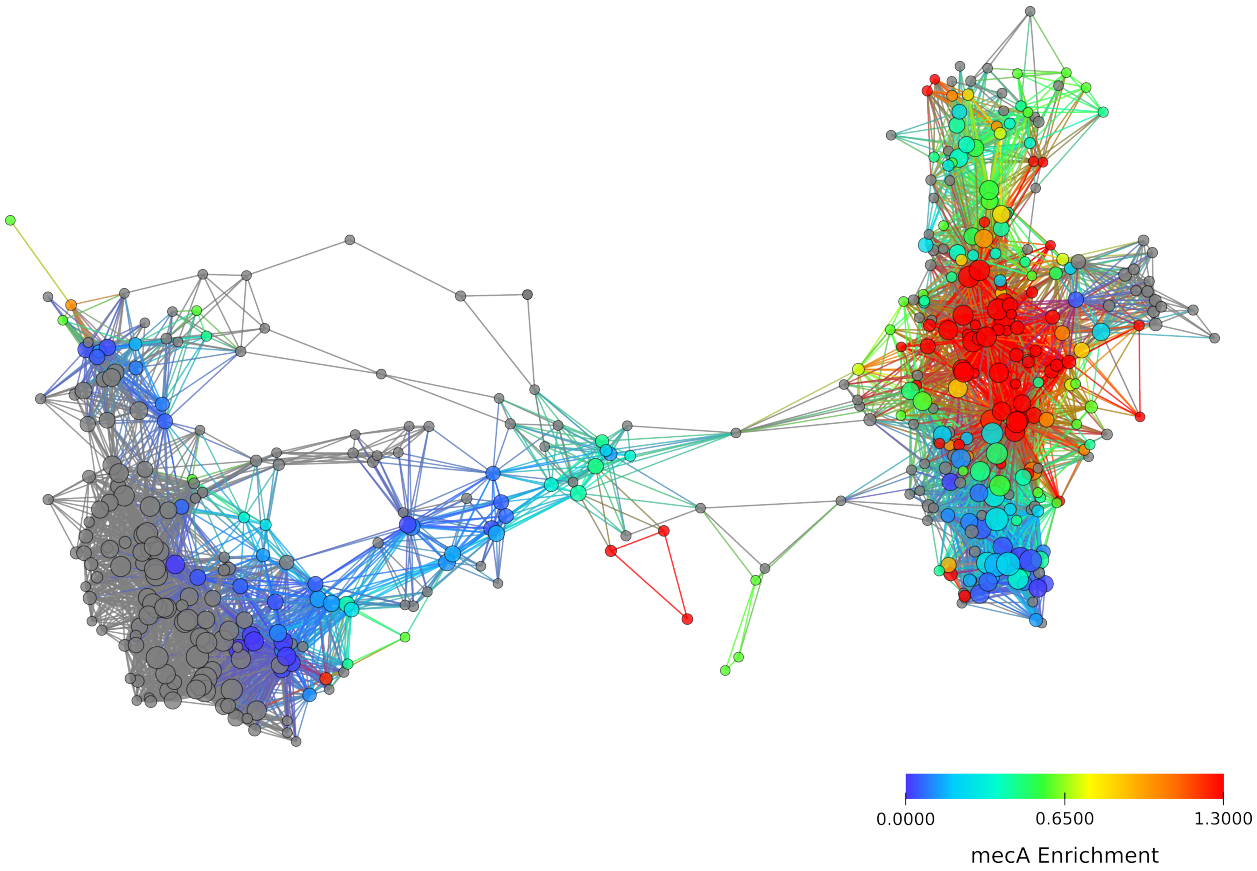


Figure 7.5: The FigFam similarity network of *S. aureus* constructed using Mapper as implemented in Ayasdi Iris [7]. We use a Hamming metric and Primary and Secondary Metric SVD filters (res: 30, gain 4x, eq.). Node color is based on strain enrichment for *mecA*, the gene conferring β-Lactam resistance. Two distinct clades of *S. aureus* are visible, one of which has already been compromised for resistance. Of important clinical significance is the growing enrichment for *mecA* in the second clade.

a similarity matrix as in previous sections, using correlation as distance. The resulting network is shown in Figure 7.6, where strains are colored by phyla-level classifications. While largely recapitulating phylogeny, the network depicts interesting correlations between phya, such as the loop between Firmicutes, Bacteroides, and Proteobacteria.

Next, we searched for genomic annotations relating to β-lactam resistance. 10 strains in the reference set had matching annotations, and we highlight those strains in the network with green diamonds. We observe resistance mostly concentrated in the Firmicutes, of which *S. aureus* is a member, however there is a strain of Proteobacteria that has acquired the resistance gene. Transfer of beta-lactam resistance into the Protebacteria is clinically

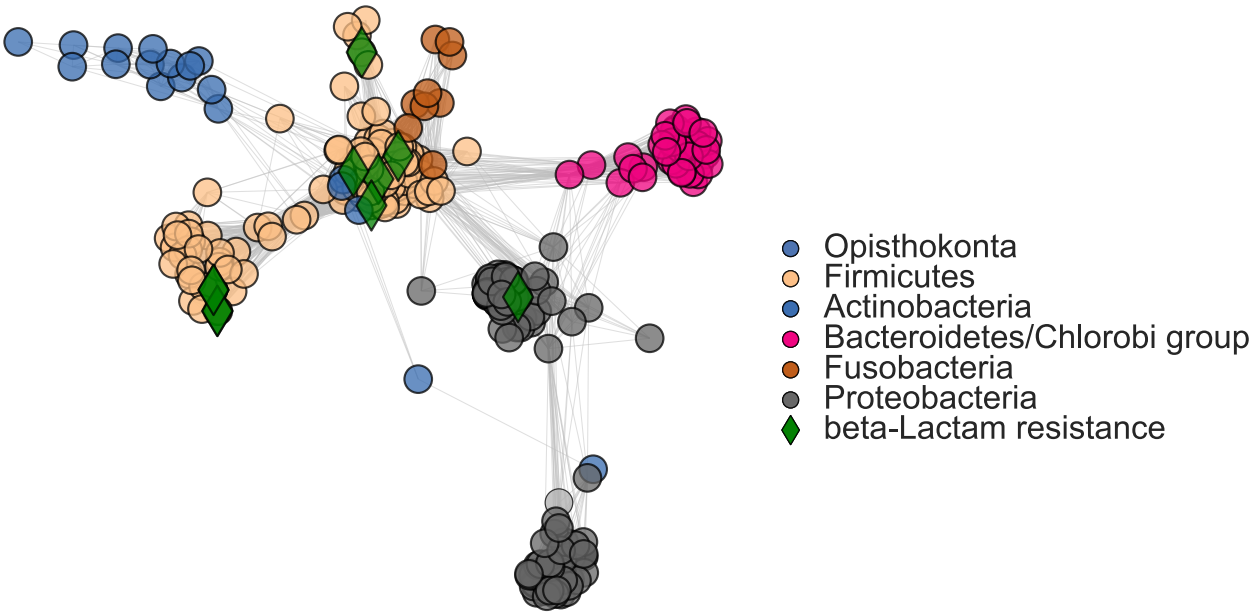


Figure 7.6: The FigFam similarity network of gastrointestinal tract reference strains identified in the Human Microbiome Project. The green diamond identifies the strains carrying resistance to β-Lactam antibiotics.

worrisome. Pathogenic proteobacteria include *S. enterica*, *V. cholerae*, and *H. pylori*, and emergence of β-lactam resistance will severely impact antibiotic drug therapies.

The species composition of each individual's microbiome can differ substantially due to a wide variety of poorly understood factors [132]. In this case, an individuals personal microbiome network will differ from the network we show in Figure 7.6, which was constructed from the set of *all* strains that have been reported across studies of multiple individuals. The relative risk for acquiring self-induced resistance will therefore vary from person to person and by the infectious strain acquired. However, a network analysis of this type will give clues as to possible routes by which antibiotic resistance may be acquired. In the clinical setting, this could assist in developing personalized antibiotic treatment regimens. We propose a more thorough expansion of this work, examining the full range of antibiotic resistance genes in order to quantify microbiome risk factors for treatment failure. We foresee an era of genomically informed infectious disease management in the clinical setting, based on an understanding of a patient's personal microbiome profile.

7.6 Conclusions

In this chapter we have used some ideas from topological data analysis to bear on problems in pathogenic microbial genetics. First, we used persistent homology to evaluate recombination rates in the core genome using MLST profile data. We showed that different pathogens have different recombination rates. We expanded this to gene transfer across the whole genome by using protein family annotations in the PATRIC database. We found different scales of recombination in different pathogens. Second, we explored the spread of MRSA in *S. aureus* populations using topological methods. We noted increasing resistance in a previously isolated population. Finally, we studied the emergence of β-lactam resistance in the microbiome, and proposed methods by which personal risk could be assessed by microbiome typing. These results point to a role for graph mining and topological data mining in health and personalized medicine.

Chapter 8

Human Recombination Rate Mapping

In this chapter, we use data from large-scale consortiums such as HapMap and the 1000 genomes project to estimate human recombination rates.

8.1 Introduction

8.2 Material

8.3 Results

8.4 Conclusions

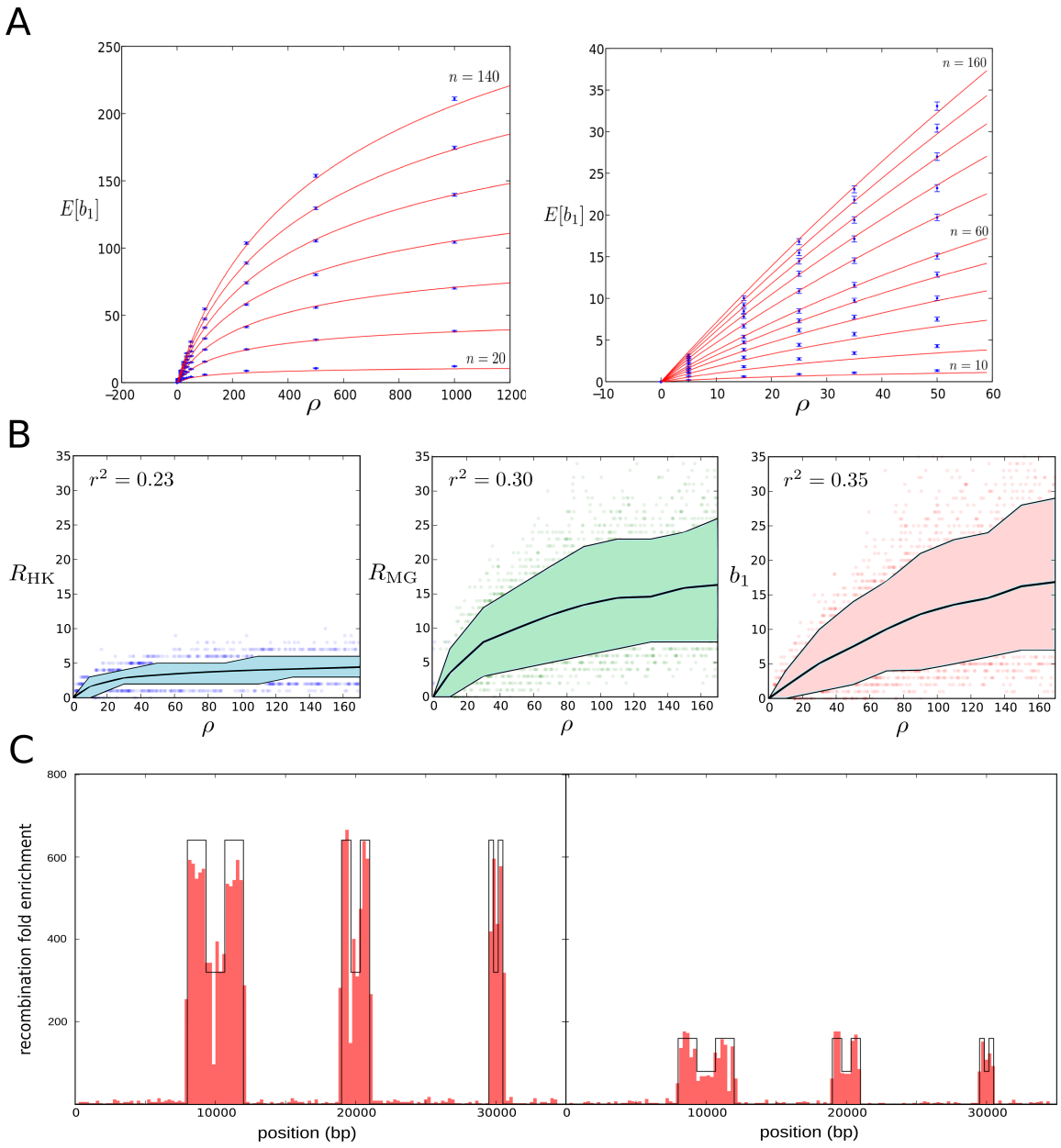


Figure 8.1: Calibration

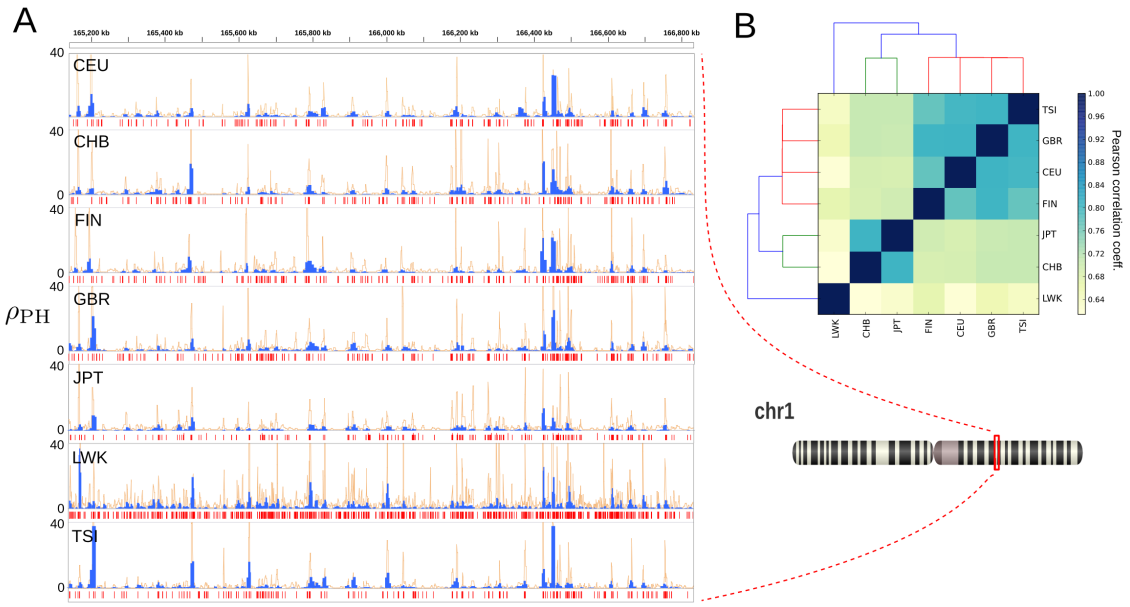


Figure 8.2: Population Tracks

Chapter 9

Multiscale Topology of Chromatin Folding

The three dimensional structure of DNA in the nucleus (chromatin) plays an important role in many cellular processes. Recent experimental advances have led to high-throughput methods of capturing information about chromatin conformation on genome-wide scales. New models are needed to quantitatively interpret this data at a global scale. In this chapter introduce the use of tools from topological data analysis to study chromatin conformation. We use persistent homology to identify and characterize conserved loops and voids in contact map data and identify scales of interaction. We demonstrate the utility of the approach on simulated data and then look data from both a bacterial genome and a human cell line. We identify substantial multiscale topology in these datasets.

9.1 Introduction

The 6 billion bases in the human genome would span a length of almost two meters if stretched end to end, yet occupy a compacted volume inside the nucleus of only a few μm^3. Even more remarkably, this million-fold level of compression is not random, but exhibits a complex hierarchical structure that intimately effects genome function through regulation of

gene expression. This multiscale pattern ranges from nucleosomes every 150 bases, promoter interactions at the megabase scale, topologically associated domains at the 10 megabase scale, and finally to organization of discrete chromosomes [42]. Chromatin conformation is dynamic, and will change throughout the cellular cycle, under the infuence of a diverse range of chromatin remodeling proteins, such as CTCF. Chromatin architecture can further be controlled epigenetically through post-translational modifications including methylation and phosphorylation.

Recently developed experimental approaches have provided unprecedented high-throughput access into the three dimensional architecture of DNA inside the nucleus [92, 42, 6]. These techniques, known as *chromosome conformation capture* (3C), use next-generation sequencing to probe for enriched physical proximity between nonadjacent genomic loci. Hi-C couples 3C with ultra-deep sequencing to measure genome-wide interaction patterns in an unbiased manner. However, while chromatin may fold in three dimensions, Hi-C contact data is only an indirect representation of these spatial relationships. Several approaches have been developed to use contact map information to generate 3D embeddings of chromatin, however this introduces additional uncertainty in the analysis [6]. Further, the contact map is an average over an ensemble of configurations. We would therefore like to directly characterize topological properties of the ensemble without the need for such an embedding.

Topological data analysis (TDA) has been applied to several problems in genomics [29, 52]. In this chapter we introduce the use of TDA to characterize the complex structure of chromatin inside the nucleus. Our primary tool is persistent homology, which extracts global information about geometric and topological invariants in data.

We first demonstrate the approach on data from simulated polymer folding. We then consider data from *C. crescentus*, a circular bacterial genome. Finally, we apply our approach to human cell line data, showing how persistent homology can capture complex multiscale folding patterns. As we show, tools from topological data analysis may prove powerful at analyzing chromatin interaction data.

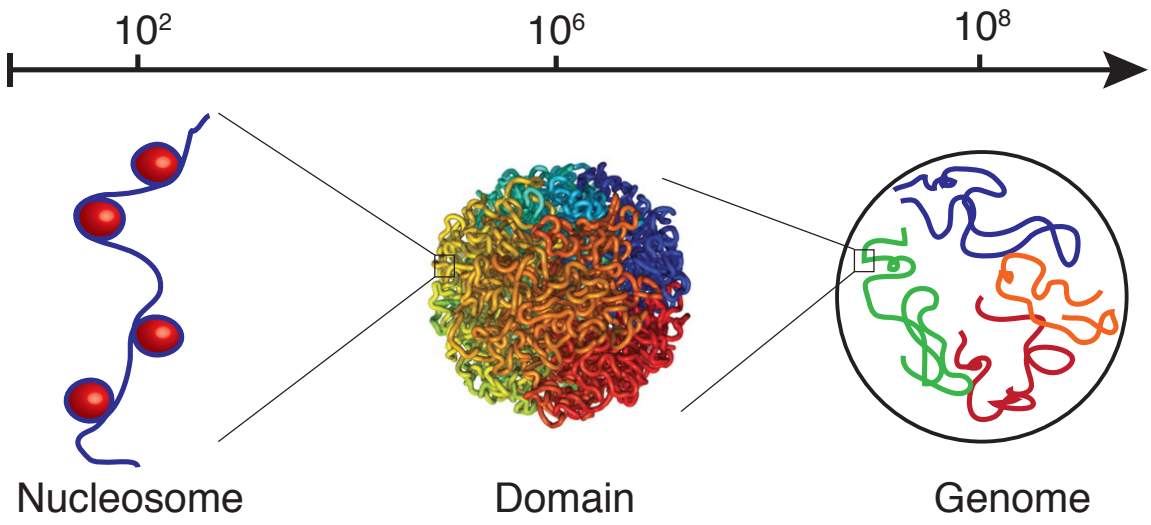


Figure 9.1: Three hierarchices of chromatin organization. At the 100 bp scale, DNA chains wrap around protein complexes called nucleosomes. At the megabase scale, these chains are compacted into domains. Lieberman-Aiden *et al.* proposed closed domains form a *fractal globule* structure. At the genome scale, chromosomes fold into the nucleus in separate territories. The fractal globule represents densely packed regions not open for transcription. Fractal globule image from [92]. Reprinted with permission from AAAS.

9.2 Background

Hi-C contact data is generated as follows: First, DNA is cross-linked in formaldehyde, linking segments of chromatin that are close in spatial proximity. This step links pieces of chromatin that are in spatial proximity. Second, pieces are fragmented and ligated to form closed loops. Finally, pieces are sheared and sequenced, and the ends of each read are mapped to loci on the genome. The data is summarized as a contact map representing counts of interactions between nonadjacent loci. For more details, see [42]. From raw frequency data, normalization procedures are then applied to account for biases. Normalization is a difficult problem and several such methods have been developed, see [6] for discussion. In this work we largely use normalized contact matrices as input. Pearson correlation ρ measures similarity between loci, which we convert to a distance as $d = 1 - \rho$.

Hi-C experiments have identified topologically associated domains. Existing computational analyses have focused on identifying significant off-diagonal contacts and associating

them with specific genomic interactions. Here we focus on the global scales of chromatin folding.

We use persistent homology to analyze Hi-C contact maps. Persistent homology captures information about loops and voids in a dataset using homology. Homology information is tracked across a scale parameter ϵ via a series of nested simplicial complexes (see [26] for more details). Invariants are summarized in a barcode diagram indexed by homology dimension H_d. Each bar in the diagram, indexed as PH_i, is annotated with a birth time, b_i, and a death time, d_i. H_1 gives information about looping between loci, and H_2 gives information about voids. Following [98], we define the *size* of a PH class as

$$x_i = \frac{b_i + d_i}{2}. \tag{9.1}$$

The distribution of PH class sizes reflects the scales of folding observed. We use Dionysus to compute persistent homology [103].

9.2.1 Long-Range Chromatin Interactions

Long-range chromatin interactions can manifest in a number of different biological consequences at megabase scales. In Figure 9.2 we show a cartoon of two possible types of interaction. On the left we see a single interaction mediated by a binding protein that brings two nonadjacent loci into contact. This could reflect a promoter-enhancer interaction, for example. We call this interaction a *one-jump loop*. On the right, we see a more complex interaction representing multiple nonadjacent loci surrounding a dense compartment of polymerase proteins. This phenomenon is known as a transcription factory, and genes adjacent to a given factory will have correlated levels of expression. We call this interaction a *multi-jump loop*, because the minimal hole generated by the filtration will span multiple nonadajacent loci.

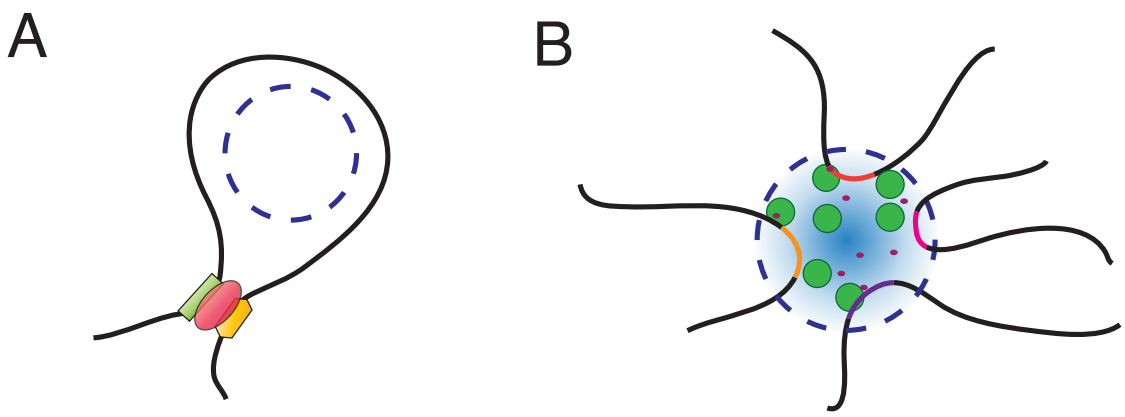


Figure 9.2: Two examples of long range chromatin interactions resulting in a topological loop. (A) A protein mediated (red) point-interaction between an enhancer (green) and promoter (yellow) sequence. (B) A transcription factory consists of dense RNA polymerase (green) around a structural core in which adjacent genomic loci (colored segments) will be cotranscribed. Transcription factors (purple) are shown.

9.2.2 Minimal Cycle Algorithm

It is important to be able to localize a cycle in order to annotate particular loops an interactions. To define a notion of minimal cycle corresponding to a PH class, we first use the contact map to locate an "essential edge" which a cycle must contain. To do this, the values of the heatmap are perturbed so that they are unique and there is a well-defined map from PH birth times to pairs of chromatin segments. That is, we can associate to each PH class (b_i, d_i) a unique "essential edge" that enters the filtration at the birth time b_i. We define a minimal cycle corresponding to (b_i, d_i) to be one containing the essential edge that traverses the shortest length along the genome, and is homologically independent from the minimal cycles of all classes born before b_i [122]. This does not uniquely specify a cycle, and we break ties by preferring cycles with shorter jumps. Specifically, if $x\ y\ \ldots\ z$ is homologous to $x\ x+1\ y\ \ldots\ z$ (where $x > y$) then the latter is considered better. We use a breadth-first search starting with the essential edge to locate a minimal cycle for a given PH class. Then, we shorten any jumps if possible.

9.3 Polymer Simulations

To explore the use of topological methods for analyzing chromatin data, we used code from [45] to simulate equilibrium folded polymer conformations. The model uses a Monte Carlo approach to simulate chromatin as a one-dimensional polymer chain confined to a volume and allowed to come to an equilibrium conformation. After equilibration, the 3D distance between monomers can be used as a measure of the contact frequency.

In Figure 9.3 we show the output of one such a simulation. Here, we simulated a 50 megabase chromatin segment as a chain of 1,000 15 nm monomers. Each monomer corresponds to approximately 6 nucleosomes, or 1200 bp. We inserted 10 fixed loops into the chain at random positions on the interior of the chain. These loops represent recurrent protein-mediated interactions and mimic chromatin folding patterns observed in real data. An ensemble of 5,000 conformations were generated and then averaged to yield the contact map depicted on the left. On the right, we show the output of persistent homology on the average contact map. Persistent homology recovers 10 H_1 intervals, consistent with the simulation and showing that topological information can be extracted from Hi-C-like contact maps.

9.4 Caulobacter Data

We examined interaction data from *Caulobacter crescentus* as published in [90]. *C. crescentus* has a 4MB circular genome. In that paper, chromatin interaction domains (CIDs) were identified at scales between 30 to 420 kb. The authors proposed a structural model consisting of brush-like plectonemes arranged along the circular fiber.

Here we look at look at sample GSM1120446, a wildtype *Caulobacter* cell. In Figure 9.4A we show the contact map data binned at 10 kb resolution. Clearly identifiable are the strong interactions along the diagonal, as well as the circular off-diagonal interactions. In Figure 9.4B is the barcode diagram computed from this contact map. Finally, in Figure 9.4C

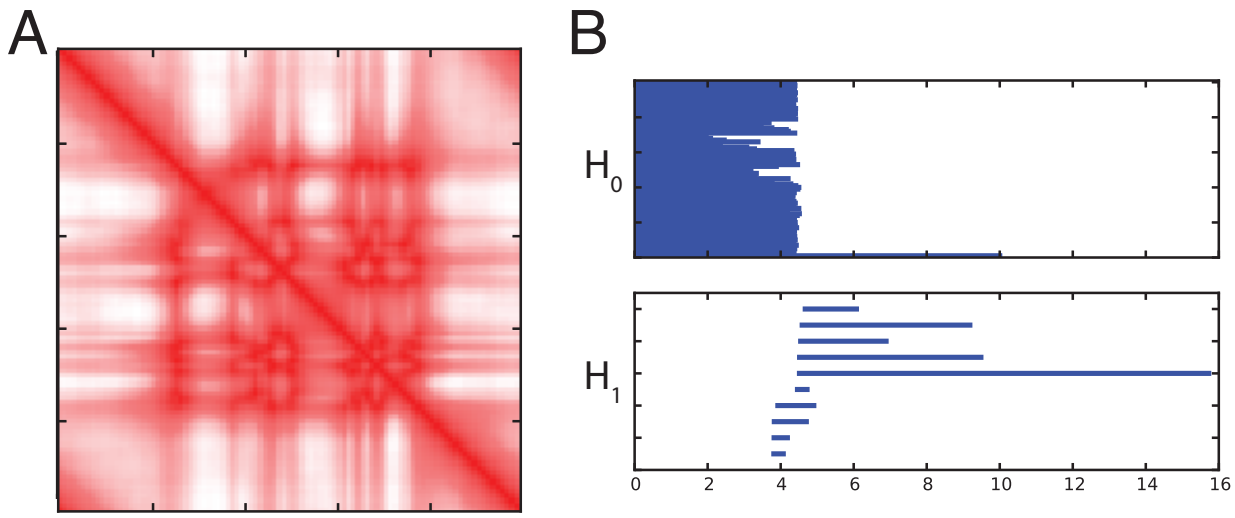


Figure 9.3: Polymer Simulation. (A) 50 Mb polymer with 10 fixed loops is allowed to reach an equilbrium conformation. (B) PH identifies 10 H_1 loops.

we see that the size of H_1 invariants is strongly bimodally distributed.

We used the minimal-cycle algorithm to determine a representative basis for each H_1 loop. Figure 9.5 shows the set of minimal cycles arranged along the genomic axis. We divide the loops between small- and large-scale loops as identified in Figure 9.4C. On the left, we see that the small-scale loops cover small genomic scales and are regularly distributed along the genome. These small loops may associate to small nucleoid-associated proteins or structural maintenance complexes (see [138]), or may simply reflect stochastic folding. On the right, we see the large-scale loops cover broader genomic regions (average size 100kb). These loops do not associate with the CIDs identified in [90], but rather reflect larger scale folding patterns.

9.5 Human Data

We examined one of the original human Hi-C data sets as published in [92]. In that paper, the authors proposed a two-compartment model of chromosomal organization associated with open and closed chromatin states and correlated with gene expression patterns. At

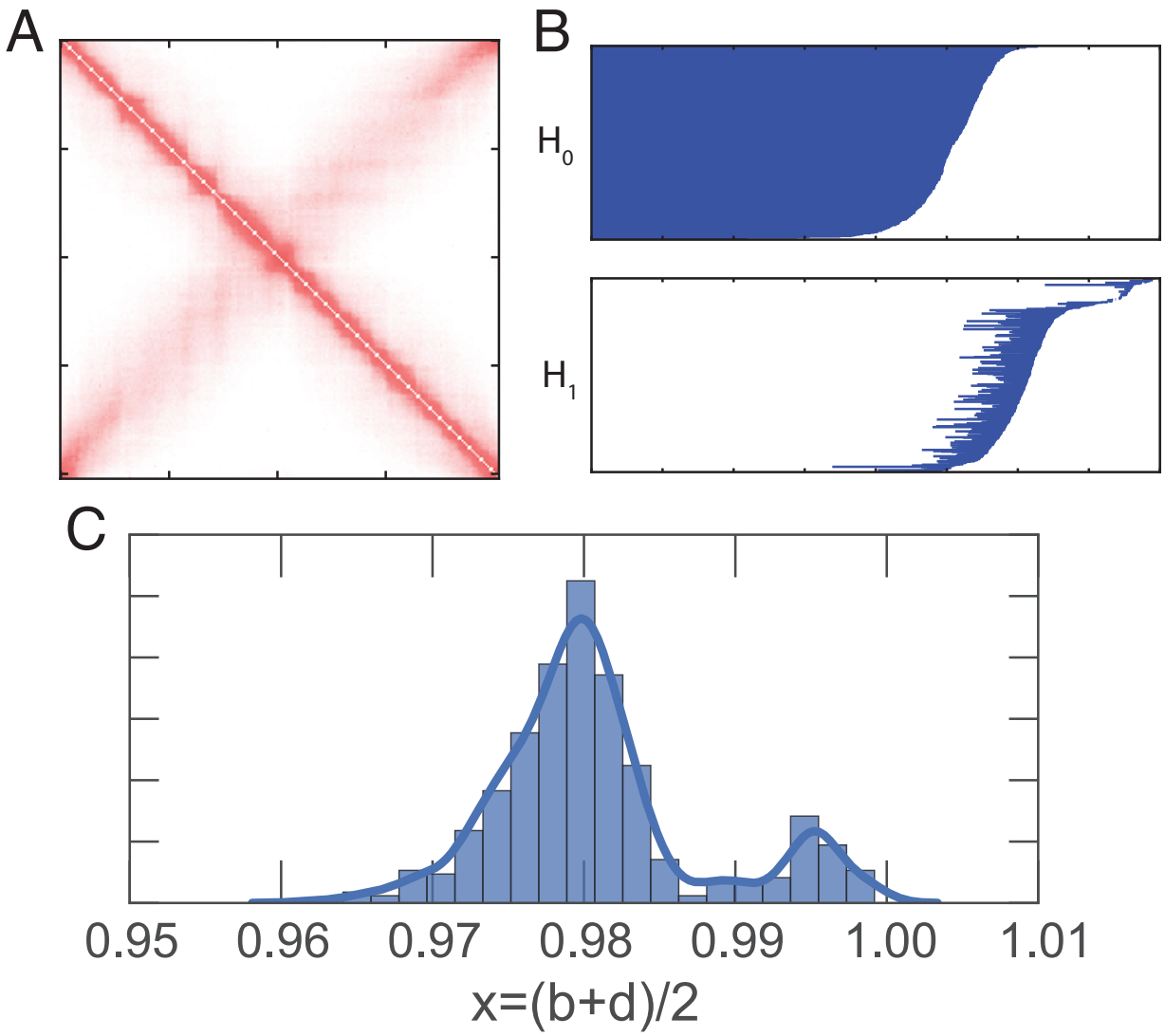


Figure 9.4: (A) Contact map and (B) barcode diagram for *Caulobacter*. (C) Distribution of H_1 bar sizes for *Caulobacter* shows a bimodal scale of folding patterns.

the megabase scale, they proposed a fractal globule model in which nearby loci along the polymer are spatially proximate in 3D.

We looked at data from GM06690, a healthy human lymphoblastoid cell line. In Figure 9.6 we show an example from chromosome 1 measured at 1 MB resolution. On the left is the observed contact map. The gray band in the middle represents the position of the centromeres. On the right is the barcode diagram computed persistent homology. We observe substantial structure in both H_1 and H_2.

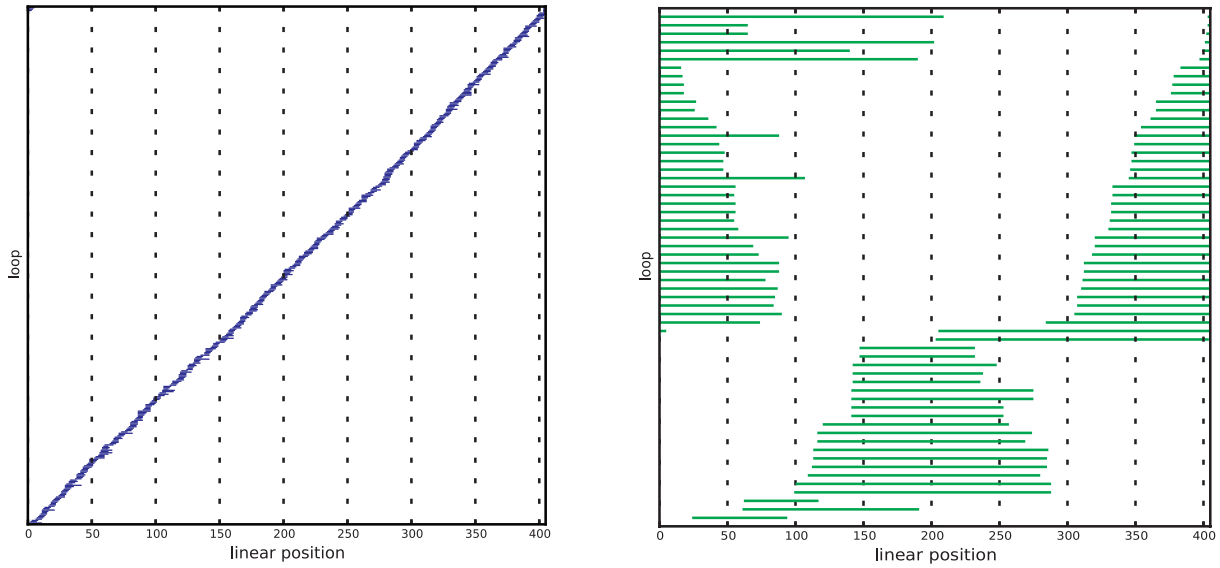


Figure 9.5: Minimal cycles projected linearly for *Caulobacter*. Left: small loops distribute uniformly across the genome. Right: pattern of large loops, which segregate into two chromosomal domains.

In Figure 9.7 we show the distribution of H_1 bar sizes. We observe a strong bimodal structure, representative of two scales of chromatin folding. This is consistent with the results in [92], which identified topologically associated domains at the 10MB scale.

Because the contact map is at 1 MB resolution, it is too coarse to capture nucleosome-level folding patterns (200 bp). More recent work has yielded Hi-C datasets at kilobase resolution [78, 116], however at this resolution the contact map is too large for an efficient persistent homology computation across the entire genome (or even an entire chromosome). It is possible that the linear nature of the chain may make a heuristic homology computation feasible.

9.6 Conclusions

Patterns of chromatin conformation inside the nucleus exhibit complex, multiscale structures that are intimately tied to genome function. Here we have used methods from TDA to characterize the scale and conformation of these structures. TDA is a natural framework

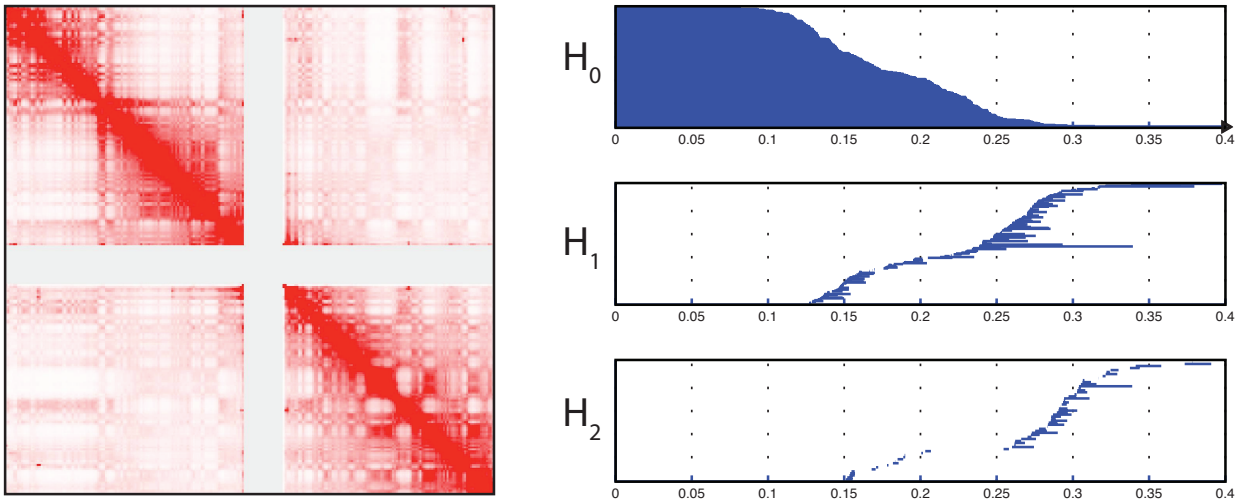


Figure 9.6: Hi-C data chromosome 1 from GM06690 human cell line data, from [92]. Left: Contact map representation. Right: PH identifies complex multiscale topology.

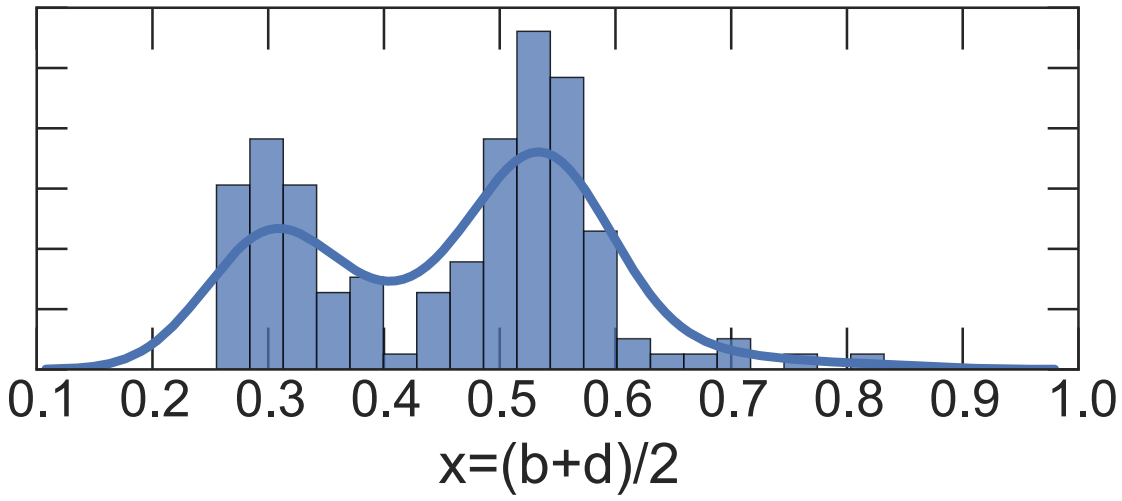


Figure 9.7: Distribution of H_1 bar sizes for GM06690 human cell line data shows a bimodal scale of folding patterns.

to study data of this type because there is a clear definition of the ambient embedding space. Using simulation we showed that persistent homology captures recurrent loops. In real data, we observed multiscale structures reflecting hierarchical patterns of chromatin organization. In the present work we have examined only intrachromosomal interactions. Interchromsomal interactions have also been reported, however the resulting contact maps are too large for homology computations at sufficient resolution. Future work will focus on identifying heuristics to improve these calculations.

Chapter 10

Conclusions

In this thesis, we have primarily considered the problem of characterizing nonvertical modes of evolution in large-scale genomic data. We have drawn on methods from topological data analysis in this task. We have developed In so doing, we have developed a framework for statistical inference using persistence diagrams. In this thesis we considered several problems in genomic and evolution. Future work will continue in this direction.

[List of things we can work on in the future.] Some other salient comments...

Need to develop more modeling. What essentiality do models have

Some concluding remarks about when persistent homology is useful. Need to understand what higher homology is telling you.

Bibliography

[1] R. Adler, O. Bobrowski, M. Borman, E. Subag, and S. Weinberger, "Persistent homology for random fields and complexes," *ArXiv.org*, 2010. arXiv: 1003.1001.

[2] M. N. Alekshun and S. B. Levy, "Molecular mechanisms of antibacterial multidrug resistance," *Cell*, vol. 128, no. 6, pp. 1037–1050, Mar. 2007. DOI: 10.1016/j.cell.2007.03.004.

[3] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zheng, W. Miller, and L. D. J, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997. DOI: 10.1093/nar/25.17.3389.

[4] M. L. Arnold, *Natural Hybridization and Evolution*, ser. Oxford Series in Ecology and Evolution. Oxford, UK: Oxford University Press, 1996.

[5] ——, *Evolution through Genetic Exchange*. Oxford, UK: Oxford University Press, 2007.

[6] F. Ay and W. S. Noble, "Analysis methods for studying the 3D architecture of the genome," *Genome Biology*, vol. 16, no. 1, p. 1306, 2015. DOI: 10.1186/s13059-015-0745-7.

[7] Ayasdi Inc., *Ayasdi Core*, 2015. [Online]. Available: http://www.ayasdi.com.

[8] H.-J. Bandelt and A. W. Dress, "A canonical decomposition theory for metrics on a finite set," *Advances in Mathematics*, vol. 92, no. 1, 1992. DOI: 10.1016/0001-8708(92)90061-o.

[9] H.-J. Bandelt, P. Forster, and A. Röhl, "Median-joining networks for inferring intraspecific phylogenies," *Molecular Biology and Evolution*, vol. 16, no. 1, pp. 37–48, 1999. DOI: 10.1093/oxfordjournals.molbev.a026036.

[10] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," *Journal of Virology*, vol. 82, no. 2, pp. 596–601, Jan. 2008. DOI:

10.1128/JVI.02005-07. [Online]. Available: http://www.ncbi.nlm.nih.gov/genomes/FLU/.

[11] F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy, "PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice," *Science*, vol. 327, no. 5967, pp. 836–840, Feb. 2010. DOI: 10.1126/science.1183439.

[12] U. Bauer, M. Kerber, and J. Reininghaus, *DIPHA (a distributed persistent homology algorithm)*, version 2.1.0, 2014. [Online]. Available: https://github.com/DIPHA/dipha/.

[13] ——, "Distributed computation of persistent homology," in *Proceedings of the Sixteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, Society for Industrial & Applied Mathematics (SIAM), 2014, pp. 31–38. DOI: 10.1137/1.9781611973198.4.

[14] U. Bauer, M. Kerber, J. Reininghaus, and H. Wagner, *Phat: Persistent Homology Algorithms Toolbox*, version 1.4.0, 2015. [Online]. Available: https://bitbucket.org/phat-code/phat.

[15] L. J. Billera, S. P. Holmes, and K. Vogtmann, "Geometry of the space of phylogenetic trees," *Advances in Applied Mathematics*, vol. 27, no. 4, pp. 733–767, 2001. DOI: 10.1006/aama.2001.0759.

[16] A. J. Blumberg, I. Gal, M. A. Mandell, and M. Pancia, "Robust statistics, hypothesis testing, and confidence intervals for persistent homology on metric measure spaces," *Foundations of Computational Mathematics*, vol. 14, no. 4, pp. 745–789, 2014. DOI: 10.1007/s10208-014-9201-4.

[17] K. Borsuk, "On the imbedding of systems of compacta in simplicial complexes," *Fundamenta Mathematicae*, vol. 35, no. 1, pp. 217–234, 1948. [Online]. Available: http://eudml.org/doc/213158.

[18] P. J. Bowler, *Evolution: The History of an Idea*. Berkeley, CA: University of California Press, 2003.

[19] G. R. Bowman, X. Huang, Y. Yao, J. Sun, G. Carlsson, L. J. Guibas, and V. S. Pande, "Structural insight into RNA hairpin folding intermediates," *Journal of the American Chemical Society*, vol. 130, no. 30, pp. 9676–9678, 2008. DOI: 10.1021/ja8032857.

[20] L. Bren, "Bacteria-eating virus approved as food additive," *FDA consumer*, vol. 41, no. 1, pp. 20–22, Jan. 2007.

[21] P. Bubenik, "Statistical topological data analysis using persistence landscapes," *Journal of Machine Learning Research*, vol. 16, pp. 77–102, 2015. [Online]. Available: http://www.jmlr.org/papers/v16/bubenik15a.html.

[22] P. Bubenik and P. T. Kim, "A statistical approach to persistent homology," *Homology, Homotopy and Applications*, vol. 9, no. 2, pp. 337–362, 2007. DOI: 10.4310/hha.2007.v9.n2.a12.

[23] D. Burke, "Recombination in HIV: An important viral evolutionary strategy," *Emerging Infectious Diseases*, vol. 3, no. 3, pp. 253–259, Sep. 1997. DOI: 10.3201/eid0303.970301.

[24] P. Camara, D. Rosenbloom, K. Emmett, A. Levine, and R. Rabadan, "Fine-scale resolution of human recombination using topological data analysis," *Cell Systems*, 2016, in press.

[25] G. Carlsson, "Topology and data," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009. DOI: 10.1090/s0273-0979-09-01249-x.

[26] ——, "Topological pattern recognition for point cloud data," *Acta Numerica*, vol. 23, pp. 289–368, 2014. DOI: 10.1017/s0962492914000051.

[27] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, "On the local behavior of spaces of natural images," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 1–12, 2008. DOI: 10.1007/s11263-007-0056-x.

[28] L. L. Cavalli-Sforza and A. W. Edwards, "Phylogenetic analysis. models and estimation procedures," *American Journal of Human Genetics*, vol. 19, no. 3, pp. 550–570, 1967. DOI: 10.2307/2406616.

[29] J. Chan, G. Carlsson, and R. Rabadan, "Topology of viral evolution," *Proceedings of the National Academy of Sciences*, vol. 110, no. 46, pp. 18 566–18 571, Nov. 2013. DOI: 10.1073/pnas.1313480110.

[30] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot, "Gromov-Hausdorff stable signatures for shapes using persistence," in *Computer Graphics Forum*, Wiley Online Library, 2009, pp. 1393–1403. DOI: 10.1111/j.1467-8659.2009.01516.x.

[31] F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman, "Subsampling methods for persistent homology," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 2143–2151. [Online]. Available: http://jmlr.org/proceedings/papers/v37/chazal15.html.

[32] F. Chazal, M. Glisse, C. Labruére, and B. Michel, "Convergence rates for persistence diagram estimation in topological data analysis," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 163–171. [Online]. Available: http://jmlr.org/proceedings/papers/v32/chazal14.html.

[33] Y. Chen, W. Liang, S. Yang, *et al.*, "Human infections with the emerging avian influenza a h7n9 virus from wet market poultry: Clinical analysis and characterisation of viral genome," *The Lancet*, vol. 381, no. 9881, pp. 1916–1925, Jun. 2013. DOI: 10.1016/S0140-6736(13)60903-4.

[34] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life," *Science*, vol. 311, no. 5765, pp. 1283–1287, 2006. DOI: 10.1126/science.1123061.

[35] J. E. Cooper and E. J. Feil, "The phylogeny of *Staphylococcus aureus* - which genes make the best intra-species markers?" *Microbiology*, vol. 152, no. 5, pp. 1297–1305, 2006. DOI: 10.1099/mic.0.28620-0.

[36] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970. DOI: 10.1038/227561a0.

[37] T. Dagan and W. Martin, "The tree of one percent," *Genome Biology*, vol. 7, no. 10, p. 118, 2006. DOI: 10.1186/gb-2006-7-10-118.

[38] C. Darwin, *On the Origin of Species by Means of Natural Selection*. London, UK: Murray, 1859.

[39] J. Davies and D. Davies, "Origins and evolution of antibiotic resistance," *Microbiology and Molecular Biology Reviews*, vol. 74, no. 3, pp. 417–433, Aug. 2010. DOI: 10.1128/mmbr.00016-10.

[40] V. de Silva and G. Carlsson, "Topological estimation using witness complexes," in *Proceedings of the First Eurographics conference on Point-Based Graphics*, Eurographics Association, 2004, pp. 157–166. DOI: 10.2312/SPBG/SPBG04/157-166.

[41] M. Deghorain and L. Van Melderen, "The staphylococci phages family: an overview," *Viruses*, vol. 4, no. 12, pp. 3316–3335, 2012. DOI: 10.3390/v4123316.

[42] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, "Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data," *Nature Reviews Genetics*, vol. 14, pp. 390–403, 2013. DOI: 10.1038/nrg3454.

[43] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science*, vol. 284, no. 5423, pp. 2124–2128, 1999. DOI: 10.1126/science.284.5423.2124.

[44] W. F. Doolittle and R. T. Papke, "Genomics and the bacterial species problem," *Genome Biology*, vol. 7, no. 9, p. 116, 2006. DOI: 10.1186/gb-2006-7-9-116.

[45] B. Doyle, G. Fudenberg, M. Imakaev, and L. A. Mirny, "Chromatin loops as allosteric modulators of enhancer-promoter interactions," *PLoS Computational Biology*, vol. 10, no. 10, e1003867, 2014. DOI: 10.1371/journal.pcbi.1003867.

[46] A. Dress, K. T. Huber, J. Koolen, V. Moulton, and A. Spillner, *Basic phylogenetic combinatorics*. Cambridge, UK: Cambridge University Press, 2011.

[47] A. Dress, K. Huber, and V. Moulton, "Some variations on a theme by buneman," *Annals of Combinatorics*, vol. 1, no. 1, pp. 339–352, 1997. DOI: 10.1007/bf02558485.

[48] A. Dress, V. Moulton, and W. Terhalle, "T-theory: An overview," *Vaccine*, vol. 17, no. 2-3, pp. 161–175, Feb. 1996. DOI: 10.1006/eujc.1996.0015.

[49] A. Dress and W. Terhalle, "The tree of life and other affine buildings," *Documenta Mathematica*, pp. 565–574, 1998. [Online]. Available: http://eudml.org/doc/233296.

[50] V. G. Dugan, R. Chen, D. J. Spiro, *et al.*, "The evolutionary genetics and emergence of avian influenza viruses in wild birds," *PLoS Pathogens*, vol. 4, no. 5, e1000076, May 2008. DOI: 10.1371/journal.ppat.1000076.

[51] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction*. American Mathematical Society, 2010. DOI: 10.1090/mbk/069.

[52] K. J. Emmett and R. Rabadan, "Characterizing scales of genetic recombination and antibiotic resistance in pathogenic bacteria using topological data analysis," in *Brain Informatics and Health*, ser. Lecture Notes in Computer Science, D. Slezak, A.-H. Tan, J. F. Peters, and L. Schwabe, Eds., vol. 8609, Springer, 2014, pp. 540–551. DOI: 10.1007/978-3-319-09891-3_49.

[53] K. Emmett and R. Rabadan, "Quantifying reticulation in phylogenetic complexes using homology," in *BICT 2015 Special Track on Topology-driven bio-inspired methods and models for complex systems (TOPDRIM4BIO)*, 2015.

[54] K. Emmett, D. Rosenbloom, P. Camara, and R. Rabadan, "Parametric inference using persistence diagrams: A case study in population genetics," in *ICML Workshop on Topological Methods in Machine Learning*, 2014.

[55] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, Apr. 2002. DOI: 10.1093/nar/30.7.1575.

[56] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh, "Confidence sets for persistence diagrams," *Ann. Statist.*, vol. 42, no. 6, pp. 2301–2339, DOI: 10.1214/14-AOS1252.

[57] B. Fasy, J. Kim, F. Lecci, C. Maria, and V. Rouvreau, *Tda: Statistical tools for topological data analysis*, version 1.4.1, 2015. [Online]. Available: https://cran.r-project.org/web/packages/TDA/index.html.

[58] J. Felsenstein, *Inferring Phylogenies*. Sunderland, Massachusetts: Sinauer Associates, 2004, ISBN: 978-0-87893-177-4.

[59] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, vol. 155, no. 3760, pp. 279–284, 1967. DOI: 10.1126/science.155.3760.279.

[60] L. R. Foulds and R. L. Graham, "The Steiner problem in phylogeny is NP-complete," *Advances in Applied Mathematics*, vol. 3, no. 1, pp. 43–49, Mar. 1982. DOI: 10.1016/s0196-8858(82)80004-3.

[61] C. Fraser, W. P. Hanage, and B. G. Spratt, "Recombination and the nature of bacterial speciation," *Science*, vol. 315, no. 5811, pp. 476–480, 2007. DOI: 10.1126/science.1127573.

[62] B. Gärtner, "Fast and robust smallest enclosing balls," in *Algorithms-ESA99*, Springer, 1999, pp. 325–338. DOI: 10.1007/3-540-48481-7_29.

[63] R. Ghrist, "Barcodes: The persistent topology of data," *Bulletin of the American Mathematical Society*, vol. 45, no. 01, pp. 61–76, 2007. DOI: 10.1090/s0273-0979-07-01191-3.

[64] J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence, "Prokaryotic evolution in light of gene transfer," *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2226–2238, 2002. DOI: 10.1093/oxfordjournals.molbev.a004046.

[65] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," *Nature*, vol. 3, no. 9, pp. 679–687, 2005. DOI: 10.1038/nrmicro1204.

[66] N. Goldenfeld and C. Woese, "Biology's next revolution," *Nature*, vol. 445, no. 7126, pp. 369–369, Jan. 2007. DOI: 10.1038/445369a.

[67] S. J. Gould, *The Structure of Evolutionary Theory*. Cambridge, MA: Harvard University Press, 2002.

[68] M. Gromov, "Hyperbolic groups," English, in *Essays in Group Theory*, ser. Mathematical Sciences Research Institute Publications, S. Gersten, Ed., vol. 8, Springer, 1987, pp. 75–263. DOI: 10.1007/978-1-4613-9586-7_3.

[69] A. Hatcher, *Algebraic Topology*. Cambridge, UK: Cambidge University Press, 2002.

[70] C. X. Hernandez, J. M. Chan, H. Khiabanian, and R. Rabadan, "Understanding the origins of a pandemic virus," 2011. arXiv: 1104.4568v1 [q-bio.PE].

[71] E. C. Holmes, E. Ghedin, N. Miller, *et al.*, "Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses," *PLoS Biology*, vol. 3, no. 9, e300, Jul. 2005. DOI: 10.1371/journal.pbio.0030300.

[72] R. R. Hudson, "Generating samples under a Wright–Fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, no. 2, 2002. DOI: 10.1093/bioinformatics/18.2.337.

[73] R. R. Hudson and N. L. Kaplan, "Statistical properties of the number of recombination events in the history of a sample of DNA sequences," *Genetics*, vol. 111, no. 1, pp. 147–164, 1985. [Online]. Available: http://genetics.org/content/111/1/147.

[74] D. H. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge, UK: Cambridge University Press, 2010.

[75] J. Huxley, *Evolution: The Modern Synthesis*. Cambridge, MA: MIT Press, 1942.

[76] International Committee on Taxonomy of Viruses, *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, A. M. Q. King, M. J. Adams, E. B. Carstens, and E. J. Lefkowitz, Eds., ser. Immunology and Microbiology 2011. Academic Press, 2012.

[77] S. O. Jensen and B. R. Lyon, "Genetics of antimicrobial resistance in *Staphylococcus aureus*," *Future Microbiology*, vol. 4, no. 5, pp. 565–582, 2009. DOI: 10.2217/fmb.09.30.

[78] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren, "A high-resolution map of the three-dimensional chromatin interactome in human cells," *Nature*, 2013. DOI: 10.1038/nature12644.

[79] K. A. Jolley and M. C. Maiden, "Bigsdb: Scalable analysis of bacterial genome variation at the population level," *BMC Bioinformatics*, vol. 11, no. 1, p. 595, 2010. DOI: 10.1186/1471-2105-11-595.

[80] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology*, ser. Applied Mathematical Sciences. Springer, Jan. 2004, vol. 157.

[81] E. C. Keen, "Phage therapy: Concept to cure," *Frontiers in microbiology*, vol. 3, Jul. 2012. DOI: 10.3389/fmicb.2012.00238.

[82] E. V. Koonin, "Darwinian evolution in the light of genomics," *Nucleic Acids Research*, vol. 37, no. 4, pp. 1011–1034, Dec. 2008. DOI: 10.1093/nar/gkp089.

[83] E. V. Koonin and Y. I. Wolf, "Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world," *Nucleic Acids Research*, vol. 36, no. 21, pp. 6688–6719, 2008. DOI: 10.1093/nar/gkn668.

[84] M. Kreitman, "Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster," *Nature*, vol. 304, no. 5925, pp. 412–417, Aug. 1983. DOI: 10.1038/304412a0.

[85] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964. DOI: 10.1007/bf02289565.

[86] R. Kwitt, S. Huber, M. Niethammer, W. Lin, and U. Bauer, "Statistical topological data analysis - a kernel perspective," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., Curran Associates, Inc., 2015, pp. 3052–3060. [Online]. Available: http://papers.nips.cc/paper/5887-statistical-topological-data-analysis-a-kernel-perspective.

[87] G. Lami, *Mcl markov cluster*, version 0.3, 2014. [Online]. Available: https://github.com/koteth/python_mcl.

[88] E. S. Lander, L. M. Linton, B. Birren, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001. DOI: 10.1038/35057062.

[89] J. G. Lawrence, G. F. Hatfull, and R. W. Hendrix, "Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches," *Journal of Bacteriology*, vol. 184, no. 17, pp. 4891–4905, 2002. DOI: 10.1128/jb.184.17.4891-4905.2002.

[90] T. B. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub, "High-resolution mapping of the spatial organization of a bacterial chromosome," *Science*, vol. 342, no. 6159, pp. 731–734, 2013. DOI: 10.1126/science.1242059.

[91] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, and J. T. Dudley, "Identification of type 2 diabetes subgroups through topological analysis of patient similarity," *Science Translational Medicine*, vol. 7, no. 311, 311ra174, Oct. 2015. DOI: 10.1126/scitranslmed.aaa9364.

[92] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, no. 5950, pp. 289–293, 2009. DOI: 10.1126/science.1181369.

[93] G. Lima-Mendez, J. Van Helden, A. Toussaint, and R. Leplae, "Reticulate representation of evolutionary and functional relationships between phage genomes," *Molecular Biology and Evolution*, vol. 25, no. 4, pp. 762–777, 2008. DOI: 10.1093/molbev/msn023.

[94] S. E. Lindstrom, N. J. Cox, and A. Klimov, "Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957–1972: Evidence for genetic divergence and multiple reassortment events," *Virology*, vol. 328, no. 1, pp. 101–119, Oct. 2004. DOI: 10.1016/j.virol.2004.06.009.

[95] M. D. Lubeck, P. Palese, and J. L. Schulman, "Nonrandom association of parental genes in influenza A virus recombinants," *Virology*, vol. 95, no. 1, pp. 269–274, 1979. DOI: 10.1016/0042-6822(79)90430-6.

[96] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson, "Extracting insights from the shape of complex data using topology," *Scientific Reports*, vol. 3, Feb. 2013. DOI: 10.1038/srep01236.

[97] S. J. Lycett, G. Baillie, E. Coulter, *et al.*, "Estimating reassortment rates in co-circulating eurasian swine influenza viruses," *Journal of General Virology*, vol. 93, no. 11, pp. 2326–2336, Nov. 2012. DOI: 10.1099/vir.0.044503-0.

[98] R. MacPherson and B. Schweinhart, "Measuring shape with topology," *Journal of Mathematical Physics*, vol. 53, no. 7, p. 073 516, 2012. DOI: 10.1063/1.4737391.

[99] W. P. Maddison, "Gene trees in species trees," *Systematic Biology*, vol. 46, no. 3, pp. 523–536, Sep. 1997. DOI: 10.2307/2413694.

[100] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec, "The GUDHI library: Simplicial complexes and persistent homology," in *The 4th International Congress on Mathematical Software (ICMS)*, Hanyang University, Seoul, Korea, Aug. 2014. DOI: 10.1007/978-3-662-44199-2_28.

[101] F. Meyer, R. Overbeek, and A. Rodriguez, "Figfams: Yet another set of protein families," *Nucleic Acids Research*, vol. 37, no. 20, pp. 6643–6654, 2009. DOI: 10.1093/nar/gkp698.

[102] Y. Mileyko, S. Mukherjee, and J. Harer, "Probability measures on the space of persistence diagrams," *Inverse Problems*, vol. 27, no. 12, p. 124 007, 2011. DOI: 10.1088/0266-5611/27/12/124007.

[103] D. Morozov, *Dionysus: A C++ library for computing persistent homology*, 2012. [Online]. Available: http://www.mrzv.org/software/dionysus/index.html.

[104] D. Müllner and A. Babu, *Python mapper: An open-source toolchain for data exploration, analysis and visualization*, version 0.1.13, 2013. [Online]. Available: http://danifold.net/mapper.

[105] V. Nanda, *Perseus: The persistent homology software*, version 4.0, 2015. [Online]. Available: http://www.sas.upenn.edu/~vnanda/perseus.

[106] M. I. Nelson and E. C. Holmes, "The evolution of epidemic influenza," *Nature Reviews Genetics*, vol. 8, no. 3, pp. 196–205, Jan. 2007. DOI: 10.1038/nrg2053.

[107] M. I. Nelson, L. Simonsen, C. Viboud, *et al.*, "Stochastic processes are key determinants of short-term evolution in influenza A virus," *PLoS Pathogens*, vol. 2, no. 12, e125, Dec. 2006. DOI: 10.1371/journal.ppat.0020125.

[108] H. C. Neu, "The crisis in antibiotic resistance," *Science*, vol. 257, no. 5073, pp. 1064–1073, 1992. DOI: 10.1126/science.257.5073.1064.

[109] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006. DOI: 10.1073/pnas.0601602103.

[110] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 7265–7270, 2011. DOI: 10.1073/pnas.1102826108.

[111] H. Ochman, J. G. Lawrence, and E. A. Groisman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, no. 6784, pp. 299–304, 2000. DOI: 10.1038/35012500.

[112] M. A. O'Malley and E. V. Koonin, "How stands the tree of life a century and a half after the origin?" *Biology Direct*, vol. 6, no. 1, pp. 1–21, 2011. DOI: 10.1186/1745-6150-6-32.

[113] L. Pachter and B. Sturmfels, *Algebraic Statistics for Computational Biology*. Cambridge, UK: Cambridge University Press, Mar. 2005.

[114] J. Penders, E. E. Stobberingh, P. H. Savelkoul, and P. F. Wolffs, "The human microbiome as a reservoir of antimicrobial resistance," *Frontiers in microbiology*, vol. 4, 2013. DOI: 10.3389/fmicb.2013.00087.

[115] C. Proux, D. van Sinderen, J. Suarez, P. Garcia, V. Ladero, G. F. Fitzgerald, F. Desiere, and H. Brüssow, "The dilemma of phage taxonomy illustrated by comparative genomics of sfi21-like siphoviridae in lactic acid bacteria," *Journal of bacteriology*, vol. 184, no. 21, pp. 6026–6036, Nov. 2002. DOI: 10.1128/JB.184.21.6026-6036.2002.

[116] S. S. P. Rao, M. H. Huntley, N. C. Durand, *et al.*, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, pp. 1–16, 2014. DOI: 10.1016/j.cell.2014.11.021.

[117] J. Reininghaus, U. Bauer, S. Huber, and R. Kwitt, "A stable multi-scale kernel for topological machine learning," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. arXiv: 1412.6821.

[118] H. Rohde, J. Qin, Y. Cui, *et al.*, "Open-source genomic analysis of Shiga-toxin–producing *E. coli* O104:H4," *New England Journal of Medicine*, vol. 365, no. 8, pp. 718–724, 2011. DOI: 10.1056/nejmoa1107643.

[119] F. Rohwer and R. Edwards, "The phage proteomic tree: A genome-based taxonomy for phage," *Journal of Bacteriology*, vol. 184, no. 16, pp. 4529–4535, 2002. DOI: 10.1128/jb.184.16.4529-4535.2002.

[120] F. Rohwer, M. Youle, and H. Maughan, *Life in Our Phage World*. San Diego, CA: Wholon, 2014.

[121] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987. [Online]. Available: http://mbe.oxfordjournals.org/content/4/4/406.

[122] B. Schweinhart, "Statistical topology of embedded graphs," PhD thesis, Princeton University, 2015.

[123] G. Singh, F. Mémoli, and G. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3D object recognition," in *Eurographics Symposium on Point-Based Graphics*, The Eurographics Association, 2007, pp. 91–100.

[124] G. J. D. Smith, D. Vijaykrishna, J. Bahl, *et al.*, "Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic," *Nature*, vol. 459, no. 7250, pp. 1122–1125, Jun. 2009. DOI: 10.1038/nature08182.

[125] M. O. Sommer, G. M. Church, and G. Dantas, "The human microbiome harbors a diverse reservoir of antibiotic resistance genes," *Virulence*, vol. 1, no. 4, pp. 299–303, 2010. DOI: 10.4161/viru.1.4.12010.

[126] Y. S. Song and J. Hein, "Constructing minimal ancestral recombination graphs," *Journal of Computational Biology*, vol. 12, no. 2, pp. 147–169, 2005. DOI: 10.1089/cmb.2005.12.147.

[127] B. Sturmfels and J. Yu, "Classification of six-point metrics," *Electronic Journal of Combinatorics*, vol. 11, R44, 2004. [Online]. Available: http://www.combinatorics.org/ojs/index.php/eljc/article/view/v11i1r44.

[128] C. A. Suttle, "Marine viruses – major players in the global ecosystem," *Nature Reviews Microbiology*, vol. 5, no. 10, pp. 801–812, Oct. 2007. DOI: 10.1038/nrmicro1750.

[129] J. K. Taubenberger and D. M. Morens, "1918 influenza: The mother of all pandemics," *Emerging Infectious Diseases*, vol. 12, no. 1, pp. 15–22, Jan. 2006. DOI: 10.3201/eid1209.05-0979.

[130] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "JavaPlex: A research software package for persistent (co)homology," in *Proceedings of ICMS 2014*, H. Hong and C. Yap, Eds., ser. Lecture Notes in Computer Science 8592, 2014, pp. 129–136. DOI: 10.1007/978-3-662-44199-2_23.

[131] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000. DOI: 10.1126/science.290.5500.2319.

[132] The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012. DOI: 10.1038/nature11234.

[133] C. M. Thomas and K. M. Nielsen, "Mechanisms of, and barriers to, horizontal gene transfer between bacteria," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 711–721, 2005. DOI: 10.1038/nrmicro1234.

[134] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. (2012). Fréchet means for distributions of persistence diagrams. arXiv: 1206.2790v2 [math.ST].

[135] L. Van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[136] J. Wakeley, *Coalescent Theory*. Roberts & Company, 2009.

[137] K. B. Walters, K. J. Emmett, J. M. Chan, D. Meroz, A. Karasin, S. Fan, G. Neumann, N. Ben-Tal, R. Rabadan, and Y. Kawaoka, "Identification of host-specific amino acids in the influenza virus PB2 polymerase subunit using machine learning approaches," *Journal of Virology*, 2016, in preparation.

[138] X. Wang, P. M. Llopis, and D. Z. Rudner, "Organization and segregation of bacterial chromosomes," *Nature Reviews Genetics*, vol. 14, no. 3, pp. 191–203, 2013. DOI: 10.1038/nrg3375.

[139] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953. DOI: 10.1038/171737a0.

[140] A. R. Wattam, D. Abraham, O. Dalay, *et al.*, "Patric, the bacterial bioinformatics database and analysis resource," *Nucleic Acids Research*, vol. 42, no. D1, pp. D581–D591, 2013. DOI: 10.1093/nar/gkt1099.

[141] B. C. Westmoreland, W. Szybalski, and H. Ris, "Mapping of deletions and substitutions in heteroduplex dna molecules of bacteriophage lambda by electron microscopy," *Science*, vol. 163, no. 3873, pp. 1343–1348, Mar. 1969. DOI: 10.1126/science.163.3873.1343.

[142] Wikipedia, *Neighbor-joining — Wikipedia, the free encyclopedia*, 2016. [Online]. Available: https://en.wikipedia.org/wiki/Neighbor_joining.

[143] C. R. Woese, "A new biology for a new century," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 2, pp. 173–186, Jun. 2004. DOI: 10.1128/mmbr.68.2.173-186.2004.

[144] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: The primary kingdoms," *Proceedings of the National Academy of Sciences*, vol. 74, no. 11, pp. 5088–5090, 1977. DOI: 10.1073/pnas.74.11.5088.

[145] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya.," *Proceedings of the National Academy of Sciences*, vol. 87, no. 12, pp. 4576–4579, 1990. DOI: 10.1073/pnas.87.12.4576.

[146] World Health Organization, "Antimicrobial resistance: Global report on surveillance 2014," Tech. Rep., 2014. [Online]. Available: http://www.who.int/drugresistance/documents/surveillancereport/en/.

[147] ——, "Influenza (seasonal) fact sheet no. 211," 2014. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs211/en/ (visited on 12/21/2015).

[148] ——, "Cumulative number of confirmed human cases of avian influenza A(H5N1) reported to WHO," 2016. [Online]. Available: http://www.who.int/influenza/human_animal_interface/H5N1_cumulative_table_archives/en/ (visited on 12/21/2016).

[149] S. Zairis, H. Khiabanian, A. J. Blumberg, and R. Rabadan. (2014). Moduli spaces of phylogenetic trees describing tumor evolutionary patterns. arXiv: 1410.0980 [q-bio.QM].

[150] E. Zuckerkandl and L. Pauling, "Molecular disease, evolution, and genetic heterogeneity," in *Horizons in Biochemistry*, M. Kasha and B. Pullman, Eds., Academic Press, 1962, pp. 189–225.

[151] ——, "Molecules as documents of evolutionary history," *Journal of Theoretical Biology*, vol. 8, no. 2, pp. 357–366, 1965. DOI: 10.1016/0022-5193(65)90083-4.