

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

2015

© 2015

Kevin Joseph Emmett

All Rights Reserved

ABSTRACT

Statistical Topology of Reticulate Evolution

Kevin Joseph Emmett

This thesis contains results of applying methods from topological data analysis to various problems in genomics and evolution. It primarily details the use of persistent homology as a tool to measure the prevalence and scale of nonvertical evolutionary events, such as reassortments and recombinations. In so doing, various techniques are developed to extract statistical information from the topological complexes that are constructed.

Contents

Contents	i
List of Figures	v
List of Tables	vii
1 Introduction	1
1.1 Molecular Evolution and the Tree Paradigm	3
1.2 Reticulate Processes and the Universal Tree	6
1.3 Evolution as a Topological Space	8
1.4 Thesis Organization	10
2 Background	13
2.1 Biology	13
2.1.1 Evolutionary Processes	13
2.1.1.1 Genes and Genomes	14
2.1.1.2 Horizontal Gene Transfer	14
2.1.2 Mathematical Models of Evolution	14
2.1.2.1 The Wright-Fisher Model	14
2.1.2.2 The Coalescent Process	15
2.1.2.3 Metrics on Sequences	16
2.1.3 Phylogenetic Reconstruction	16
2.1.3.1 Distance Matrix Methods	17
2.1.3.2 Phylogenetic Networks	17
2.1.3.3 Space of Phylogenetic Trees	17
2.1.3.4 Number of Trees	17
2.2 Topological Data Analysis	18
2.2.1 Intuition	18
2.2.2 Constructing Spaces from Real Data	18
2.2.2.1 The Čech and Vietoris-Rips Complexes	18
2.2.2.2 Filtrations	19
2.2.3 Condensed Representations	19
2.2.3.1 The Mapper Algorithm	19
2.2.4 Persistent Homology	20
2.2.4.1 The Persistence Algorithm	21

2.2.4.2	Stability	21
2.2.4.3	Statistical Persistent Homology	23
2.2.4.4	Multidimensional Persistence	24
2.3	Applying TDA to Molecular Sequence Data	24
2.3.1	The Simplest Example	27
2.3.2	The Space of Trees, Revisited	27
I	Theory	29
3	Quantifying Reticulation Using Topological Complex Constructions	31
3.1	Introduction	31
3.2	Persistent Homology of Sequence Data	33
3.2.1	Vertical Evolution	33
3.2.2	Reticulate Evolution	34
3.3	Reticulation Quantification Using Homology	35
3.4	Examples	35
3.5	The Median Complex Construction	37
3.5.1	Inclusion	38
3.5.2	Split Decomposition	39
3.6	Interpretation of Higher Dimensional Homology	39
3.7	Čech Complex Construction as an Optimization Problem	40
3.7.1	Molecular Hypothesis	41
3.8	Examples	42
3.8.1	Kreitman Data	42
3.8.2	Buttercup Data	43
3.8.3	Additional Examples	43
3.8.4	Simple Examples	43
3.9	Conclusions	44
4	Parametric Inference using Persistence Diagrams	45
4.1	Introduction	45
4.2	Warmup: Gaussian Random Fields	47
4.3	The Coalescent Process	47
4.4	Statistical Model	48
4.5	Experiments	51
4.5.1	Coalescent Simulations	51
4.6	Conclusions	52
II	Applications: Viruses and Bacteria	53
5	Phage Mosaicism	55
5.1	Introduction	55
5.2	Approach	58

5.2.1	Data	58
5.3	Results	58
5.4	Phage Ecological Properties	59
5.5	Conclusions	59
6	Reassortment in Influenza Evolution	61
6.1	Introduction	61
6.2	Influenza Virus	61
6.3	Reassortment	63
6.4	Multiscale Flu Reassortment	64
6.5	Prediction of Host Specific Residues	67
6.6	Conclusions	67
7	Reticulate Evolution in Pathogenic Bacteria	69
7.1	Introduction	69
7.2	Evolutionary Scales of Recombination in the Core Genome	72
7.3	Protein Families as a Proxy for Genome Wide Reticulation	74
7.4	Antibiotic Resistance in <i>Staphylococcus aureus</i>	76
7.5	Microbiome as a Reservoir of Antibiotic Resistance Genes	77
7.6	Conclusions	80
8	Prokaryote Reticulate Evolution - Tree of Life	81
8.1	Introduction	81
8.2	Materials and Methods	81
8.3	Results	82
8.4	Conclusion	82
III Applications: Human Data		83
9	Human Recombination Rate Mapping	85
9.1	Introduction	85
9.2	Material	85
9.3	Results	85
9.4	Conclusions	85
10	Multiscale Topology of Chromatin Folding	89
10.1	Introduction	89
10.2	Background	91
10.2.1	Long-Range Chromatin Interactions	92
10.2.2	Minimal Cycle Algorithm	93
10.3	Polymer Simulations	94
10.4	Caulobacter Data	94
10.5	Human Data	95
10.6	Conclusions	97

11 Conclusions	101
Bibliography	103

List of Figures

1.1	Charles Darwin’s Evolutionary Tree	2
1.2	Carl Woese’s Three Kingdom Tree of Life	5
1.3	Ford Doolittle’s Reticulate Tree of Life	7
1.4	The coffee mug and the donut	10
1.5	Treelike and reticulate phylogenies	10
2.1	Dimensionality Reduction for EDA	19
2.2	A filtration	20
2.3	The Persistence Pipeline	21
2.4	Applying TDA to Sequence Data	26
3.1	A tree is trivially contractible and has vanishing higher homology.	34
4.1	Two representations of the same topological invariants computed using persistent homology	46
4.2	Distributions of statistics defined on the H_1 persistence diagram for different model parameters	49
4.3	Inference of recombination rate ρ using topological information	51
5.1	Inconsistency of morphological classification in bacteriophage.	57
5.2	Bacteriophage Barcode Diagram	58
6.1	Structure of an influenza virus particle	63
6.2	flu...	65
6.3	H_1 persistence diagram computed from an avian influenza dataset.	66
7.1	Core genome exchange in <i>K. pneumoniae</i> and <i>S. enterica</i>	73
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. <i>H. pylori</i> has an earlier scale of recombination not present in the other species.	74
7.3	Relative recombination rates computed by persistent homology from MLST profile data.	75
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.	76

7.5	FigFam similarity network of <i>S. aureus</i>	78
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.	80
9.1	Calibration	86
9.2	Population Tracks	87
10.1	Three hierarchies 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 <i>et al.</i> proposed closed domains form a <i>fractal globule</i> 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 [57]. Reprinted with permission from AAAS.	91
10.2	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.	93
10.3	Polymer Simulation. (A) 50 Mb polymer with 10 fixed loops is allowed to reach an equilibrium conformation. (B) PH identifies 10 H_1 loops.	95
10.4	(A) Contact map and (B) barcode diagram for <i>Caulobacter</i> . (C) Distribution of H_1 bar sizes for <i>Caulobacter</i> shows a bimodal scale of folding patterns.	96
10.5	Minimal cycles projected linearly for <i>Caulobacter</i> . Left: small loops distribute uniformly across the genome. Right: pattern of large loops, which segregate into two chromosomal domains.	97
10.6	Hi-C data chromosome 1 from GM06690 human cell line data, from [57]. Left: Contact map representation. Right: PH identifies complex multiscale topology.	98
10.7	Distribution of H_1 bar sizes for GM06690 human cell line data shows a bimodal scale of folding patterns.	98

List of Tables

2.1	Dictionary connecting algebraic topology and evolutionary biology	26
3.1	Čech Homology of Hypercube	42
5.1	Phage families defined by the ICTV	56
7.1	Pathogenic bacteria selected for study and sample sizes in each analysis.	71

Chapter 1

Introduction

Darwin's *On the Origin of Species* contains a single figure, depicting the ancestry of species as a branching genealogical tree [25] (see Figure 1.1). Since then, the tree structure has been the dominant framework to understand, visualize, and communicate discoveries about evolution. Indeed, a primary focus of evolutionary biology has been to expand the *universal tree of life*, the set of evolutionary relationships among all extant and extinct organisms on Earth [11]. Traditionally, this was the realm of phenotype-derived taxonomies. With the advent of molecular sequence data and computational approaches for tree-inference, molecular phylogenetics has become the standard tool for inferring evolutionary relationships. However, a tree is accurate only if the Darwinian model of descent via reproduction 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 [3]. 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

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

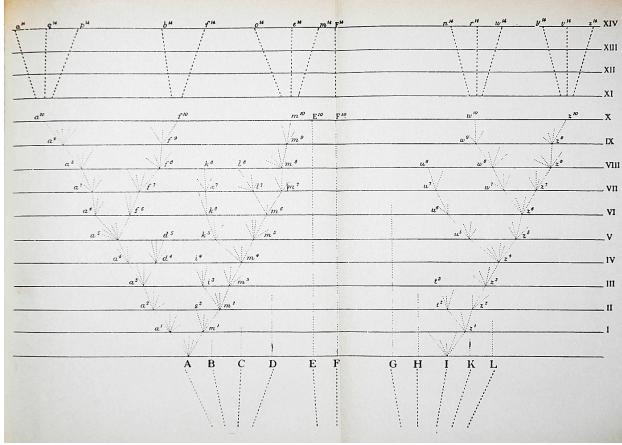


Figure 1.1: The only figure in Darwin’s *On the Origin of Species*. Darwin argued for descent with modification and natural selection as the driving processes underscoring evolution. In this figure, Darwin sketched his idea for how diverging species would result in a tree structure. Reproduced from [25].

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 [28, 69].

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 data, (2) develop approaches tailored to the unique problems and assumptions inherent in genomic data, and (3) apply these approaches to a wide range of biological datasets in which reticulate processes 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 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 [87]. 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) or small insertions and deletions of a few nucleotides (indels).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.

²See [46] and [41] for historical detail.

³In humans, only 1.5% of the genome is protein-coding, the rest largely non-functional. [53]. 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.

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

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 [91, 92]. It became apparent 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. If one has a larger set of sequenced organisms, computing the complete set of pairwise distances gives a *distance matrix* for the organisms. From the distance matrix, one attempts to associate a tree to the data such that pairwise distances along the tree are close to the pairwise sequence distances. Walter Fitch and Emanuel Margolish helped popularized this approach by constructing a weighted least squares approach to fitting phlyogenetic trees from distance matrices [36]. 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, quartet analysis, and Bayesian methods, we will focus on distance matrix methods because of their close connection with the topological ideas we employ later.

One important early result from molecular phylogenetics was Carl Woese's organization of bacteria, eukarya, and archaea into the three domains of life [89]. 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 *archaeabacteria* was more closely related to eukaryotes than were the rest of the prokaryotes. This led to the three-domain system of life (see Figure 1.2).

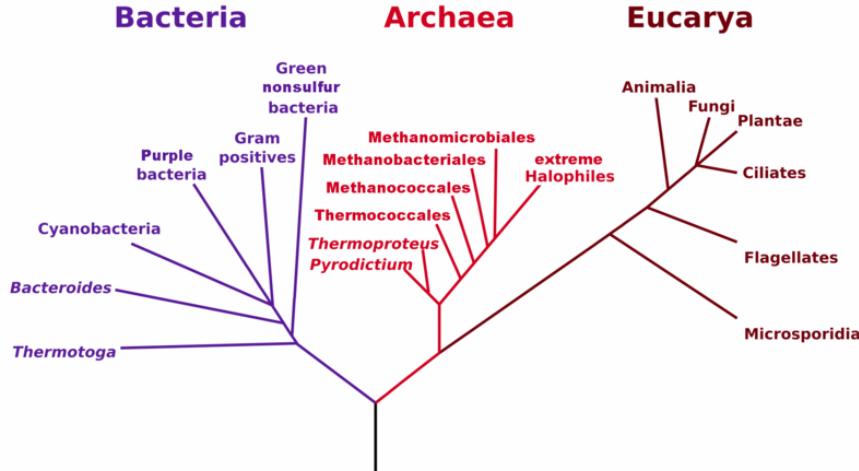


Figure 1.2: Carl Woese’s three kingdom tree of life. Using 16S subunit ribosomal RNA, Carl Woese identified archaea as a distinct phylogenetic kingdom. Previously, based on morphological similarity (specifically, unicellular and lacking a nucleus), archaea had been grouped with bacteria. This result was an early success for molecular phylogenetics and the use of conserved gene segments for molecular classification. Figure adapted from [90].

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 nineteenth century, 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 Woese's observation had, there remains a subtle difficulty, which Woese himself would come to contemplate in later work. Woese's phylogeny was based on only 1,500 nucleotides in the ribosomal RNA, less than 1% of the length of a typical bacterial genome (see [24]). Even more striking, 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 [21], 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 Second, even among universal genes, the presence of reticulate evolutionary processes will confound systematic analysis.

The model of a bifurcating tree is only consistent if. When organisms exchange genetic material by means other than direct reproduction, ancestral relationships will be

If one were to use a different genomic region to construct a tree of species, a different tree topology would be generated.

Horizontal exchange occurs when a donor bacteria transmits foreign DNA into a genetically distinct bacteria strain. Three mechanisms of horizontal transfer are identified, depending on the route by which foreign DNA is acquired [68]. 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).

Incompatibilities in the tree paradigm now appear as the rule, not the exception, demonstrating the need for new representations of evolutionary relationships [28, 29]. Many have argued that, in light of genomic evidence for the prevalence of reticulate processes, the very notion of a universal tree of life must be discarded.

One wonders if the information deduced from small genomic sections can be extrapolated to other regions, as different gene sequences can yield vastly different tree topologies. [cite]

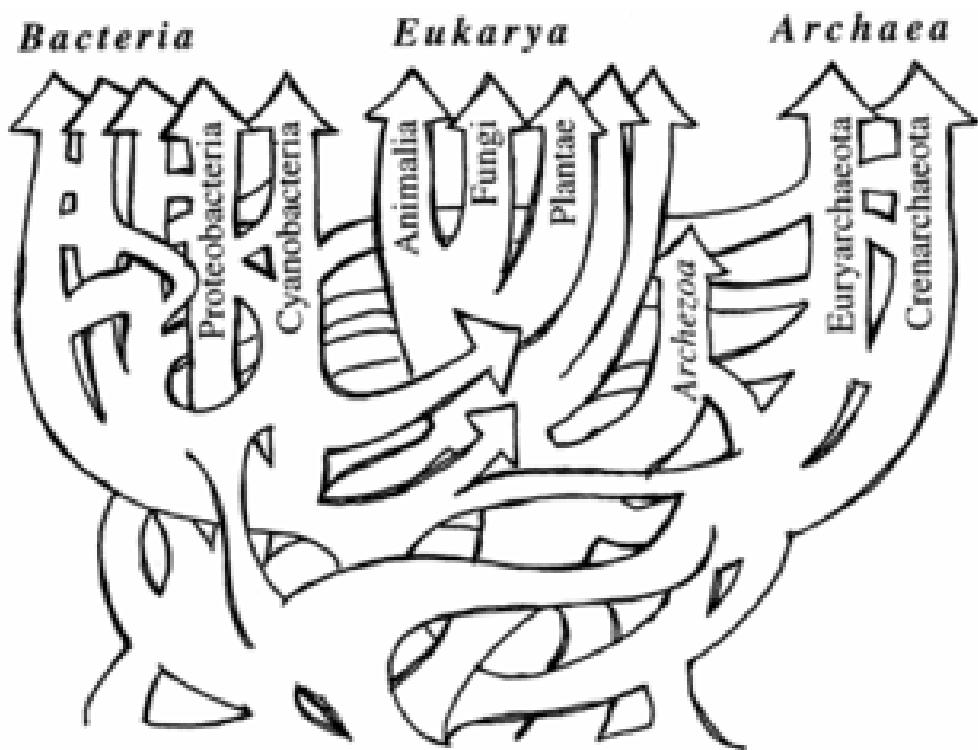


Figure 1.3: W Ford Doolittle's representation of the universal tree of life with reticulate evolution. While the three domains of life are still recognizable, patterns of divergence no longer follow a strictly treelike model. (From *Science*, vol. 284, issue 5423, page 2127. Reprinted with permission from AAAS.)

Doolittle, Koonin]. These and other similar situations, further described below, call for new methods of characterizing evolutionary relationships.

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 [14]. In influenza, segmental gene reassortments lead to antigenic novelty and the emergence of epidemics [65]. In several pathogenic bacteria, including *E. coli* and *S. aureus*, horizontal gene transfer has been responsible for the spread of antibiotic resistance genes [68, 2].

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. Here, we briefly unpack this idea.

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. Objects are considered topologically equivalent if they can be deformed into one another without introducing 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, and can be freely deformed into one another. Topologically, the two objects are equivalent.⁵

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. 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 loops in our space. Equivalently, this is the number of cuts needed to transform the space into something contractible to a point. 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. On the left we have a simple rooted three leaf evolutionary topology. There is a single connected component that is

⁵The two objects are topologically equivalent to a solid torus, which can be represented as $D^2 \times S^1$, a solid two-dimensional disk wrapping around a circle.

trivially contractible, giving $b_0 = 1$ and $b_1 = 0$. By trivially contractible we mean it can be deformed into a point, a property which will hold for all tree topologies. On the right, we have a reticulate topology again involving three leaves. We can envision the center leaf as being a reticulate offspring of parents ancestral to the left and right leaves. That is, the center leaf carries some genetic material from both the left and right branches. Accounting for this reticulation generates a single loop, giving $b_0 = 1$ and $b_1 = 1$. The object is no longer treelike, but characterized by more complex topology. The Betti numbers therefore 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 many holes being formed at points where reticulate events have occurred. 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 to construct topological spaces from finite samples, how to make comparisons among gene sets, and how to handle extinct organisms. We will address these questions, and in doing so develop new techniques to construct and extract statistical information from topological spaces. In doing so, we provide a fuller understanding of evolutionary relationships than allowed by current phylogenetic methods.

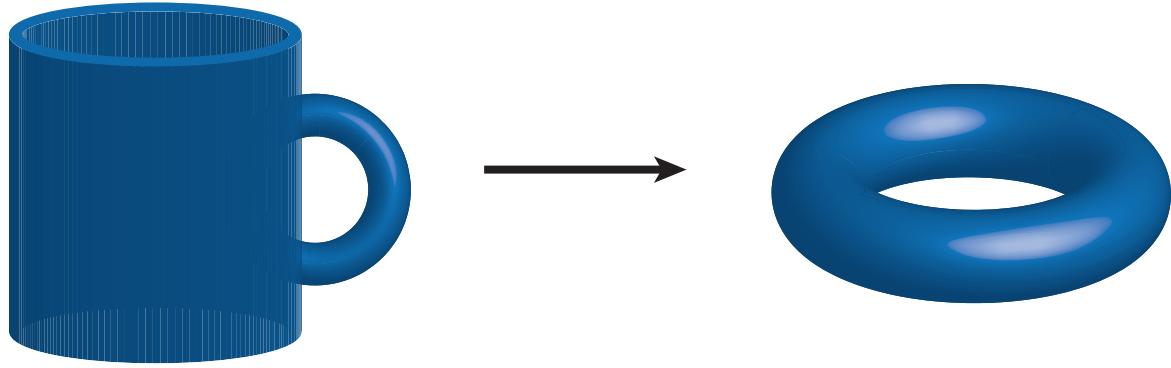


Figure 1.4: The paradigmatic topology example. The coffee mug can be continuously deformed into the donut and are therefore topologically equivalent. Both exhibit the topology of a solid torus ($D^2 \times S^1$).

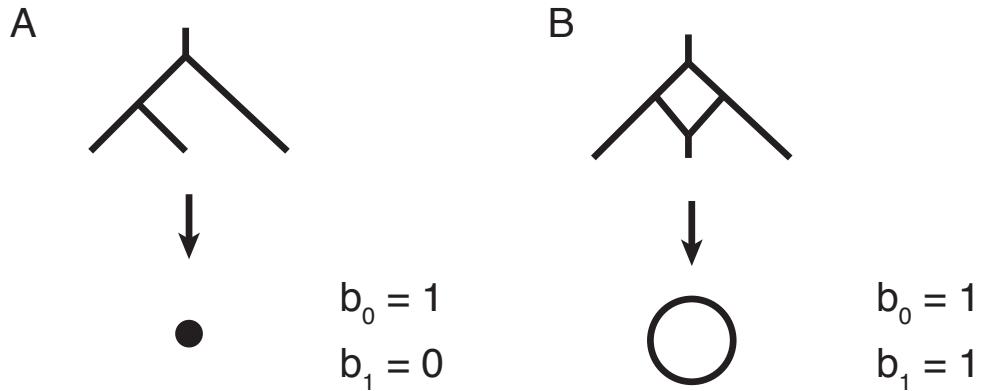


Figure 1.5: (A) A simple treelike phylogeny is contractible to a point. (B) A reticulate phylogeny that is equivalent to a circle and not contractible without a cut. The two spaces are not topologically equivalent and can be distinguished by their Betti numbers.

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 suited 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 [33]. 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 [34]

In Part ??, we apply our approach to several problems in evolution and genomics. In Chapter 5 we study bacteriophages. In Chapter ?? 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 study prokaryotic evolution and species tree topologies. In Chapter 9 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 10 we analyze Hi-C data to explore patterns of chromatin folding in the nucleus in both prokaryotic and human datasets.

Finally, in Chapter 11 we summarize these results and present future research directions.

Chapter 2

Background

This thesis integrates open problems in evolutionary biology with newly developed tools from applied topology. As few readers are likely to have substantial exposure to both fields, we use this initial chapter to supply sufficient background to motivate later discussion. Exposition required for specific results can be found in their respective individual chapters.

2.1 Biology

In this section we present biological background. This section is intended to motivate the problems discussed. More specific background relevant to the applications described in Part II and Part III is presented in those Parts.

What do we mean by horizontal and vertical evolution? Vertical, or clonal, evolution is mediated by stochastic mutations over multiple generations. Vertical evolution will be consistent with a phylogenetic tree. Horizontal evolution...

2.1.1 Evolutionary Processes

Vertical: clonal. Mutation and drift. Horizontal; reticulate; lateral. Recombination and reassortment. Gene trees vs species trees.

2.1.1.1 Genes and Genomes

The information required for biological function is contained in an organisms genome. The genome is the linear string of nucleotides that represent the set of genes in an organism. Following the fundamental dogma of biology, DNA is transcribed into RNA, RNA is translated into amino acids, and amino acids are folded into proteins. Proteins are the functional element of biology.

Beyond simply coding for an organisms function, the genome includes an imprint of the evolutionary history that gave rise to that particular organism. Comparing the genomes of multiple organisms. This is the field of *comparative genomics*.

2.1.1.2 Horizontal Gene Transfer

Discussion of nonvertical evolutionary processes. This will include references to Doolittle, Koonin, Gogarten. Lateral Gene Transfer. Species Trees and Gene Trees. Recombination and reassortment.

2.1.2 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. Large-scale genome surveys. Genealogy and population structure.

2.1.2.1 The Wright-Fisher Model

The Wright-Fischer model is a Forward time simulation of population. Here we describe the simplest Wright-Fisher model: neutral, constant population size N . At each time step randomly choose a parent. [Figure comparing Wright-Fisher and Coalescent.]

2.1.2.2 The Coalescent Process

The coalescent process is a stochastic model that generates the genealogy of individuals sampled from an evolving population [85]. 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:

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

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` [43].

2.1.2.3 Metrics on Sequences

For the sequences from different organisms to be compared, they must first be aligned. A sequence alignment is an arrangement of the characters in a set of sequences into a set of columns such that characters sharing evolutionary identity are in the same column. In this case insertion and deletions can introduce gaps. Alignment can be a complex part of any sequence analysis, particularly when there is substantial divergence between sequences of interest. Performing sequence alignment is largely beyond the scope of this thesis, and in general we assume sufficient sequence similarity such that alignment can be performed without difficulty.

What metrics can we put on aligned sequences? The simplest model, and the one most commonly adopted in this thesis, is the Hamming metric, which simply counts the differences.
Hamming metric: simple count of differences between

More biologically motivated metrics will incorporate some model of evolution and account for the possibility of back mutation. These include Jukes-Cantor, Nei-Tamura, etc.

Jukes-Cantor metric is defined as

$$d = -\frac{3}{4} \ln(1 - \frac{4}{3}p). \quad (2.2)$$

2.1.3 Phylogenetic Reconstruction

Phylogenetics is concerned with relationships among species as inferred from evolutionary characters. In practice: tree-building.

Start with sequences. Perform an alignment. From an alignment, one can then directly use parsimony or likelihood approaches. Alternatively, one can compute a matrix of pairwise distances and then construct a tree that best approximates these distances. Most relevant to this thesis are the distance-based approaches (because they can be viewed as finite metric spaces amenable to topological analysis). Only in the case of perfectly additive data will a tree be able to exactly fit the matrix. Identifying the pairwise distance matrix with a its

finite metric space representation allows most of the results described here.

Include discussion of rooted vs unrooted.

2.1.3.1 Distance Matrix Methods

Introduced by Cavalli-Sforza and Edwards in 1967 [17] and Fitch and Margoliash in 1967 [36]. Compute a matrix of pairwise distances and then find the tree that best approximates those distances. Weighted and unweighted least squares. UPGMA. Neighbor joining is now the most common distance-matrix approach because it can perfectly reconstruct an additive tree. Neighbor joining was introduced by Saitou and Nei in 1987 [76].

There are limitations of distance based methods. Do not use all the information.

Need to include discussion of metrics on aligned sequences.

2.1.3.2 Phylogenetic Networks

Several existing methods for representing reticulate evolution. Most assume the simplest generalization. However, the resulting networks can be complex and difficult to interpret quantitatively. Split Networks. (Look at Morrison Review.)

2.1.3.3 Space of Phylogenetic Trees

Studies of tree space were initiated by Billera, Holmes, and Vogtmann in [8]. Tree space on L leaves is modeled as a subspace of $\mathbb{R}^{(XX)}$, where In their model, each point represents an unrooted binary tree with L leaves and positive branch lengths. Number of interior edges $r = L - 3$, a particular additive tree can be plotted as a point in the positive open orthant $(0, \infty)^r$. A single orthant corresponds to a single tree topology. [Include treespace figure].

2.1.3.4 Number of Trees

The number of unrooted bifurcating tree topologies with L leaves is $(2L - 5)!!$. This can be most easily seen via induction. For $L = 3$ leaves there is 1 unrooted tree topology and

3 line segments. To move to the case of $L = 4$, we can add the fourth leaf to any of the 3 segments. This results in 3 possible topologies with 5 total line segments (4 external leaves and 1 internal segment).

The number of rooted bifurcating tree topologies with L leaves is $(2L - 3)!!$. The number of unrooted bifurcating tree topologies with L leaves is $(2L - 5)!!$. As can be seen, the number of tree topologies explodes with the number of leaves. Fitch quote: *more than 20 species, more than Avogadro's number of topologies.* Phylogeny can be seen as projecting onto the space of trees.

2.2 Topological Data Analysis

2.2.1 Intuition

2.2.2 Constructing Spaces from Real Data

With a metric we can represent this as a finite metric space. How can we apply the ideas of topology to these spaces?

Finite Metric Spaces Finite metric spaces are really interesting objects of study. Lots of good work on embedding finite metric spaces, see Matousek. Metric space with a finite number of points. Combinatorics, geometric, and algorithms. In topological data analysis, our spaces of interest are finite metric spaces.

2.2.2.1 The Čech and Vietoris-Rips Complexes

The Čech complex consists of the set of simplices σ with vertices $v_1, \dots, v_k \in S$ such that

$$\cap_i B(v_i, \epsilon) \neq \emptyset \quad (2.3)$$

The Vietoris-Rips complex is defined as

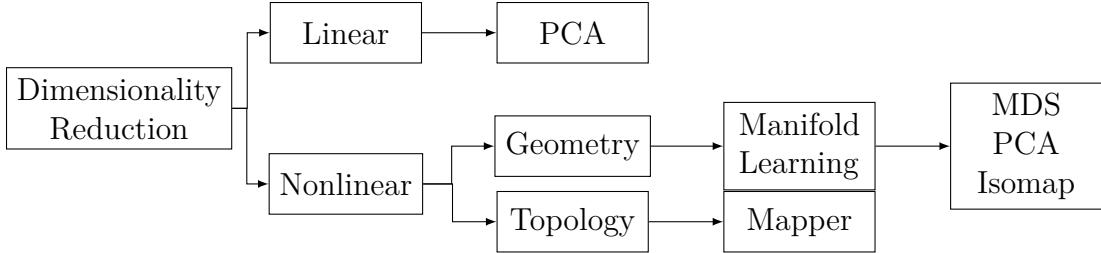


Figure 2.1: Dimensionality Reduction for EDA

$$\text{VietorisRips}(r) = \{\sigma \in S \mid \text{diam}(\sigma) \leq 2r\} \quad (2.4)$$

where $\text{diam}(\sigma) = \{\sup d(i, j) \mid i, j \in \sigma\}$

2.2.2.2 Filtrations

A series of inclusions.

2.2.3 Condensed Representations

2.2.3.1 The Mapper Algorithm

Condensed Representations. Exploratory data analysis seeks to represent high-dimensional datasets in .

Mapper algorithm was first developed in [78]. Mapper is coordinate free and depends onnly on the similarity of points as measured by the distance function. Further exposition can be found in [59]. Used in biology example in [67] top classify breast cancer subtypes. In our work we use the commercial Mapper implementation Ayasdi [5]. An open-source implementation of the Mapper algorithm is avvailable in the Python Mapper package [64].

Steps: (1) Project using filter function. (2) Create overlapping bins (3) Cluster in the projected space. (3) Connect pairs of bins with shared points

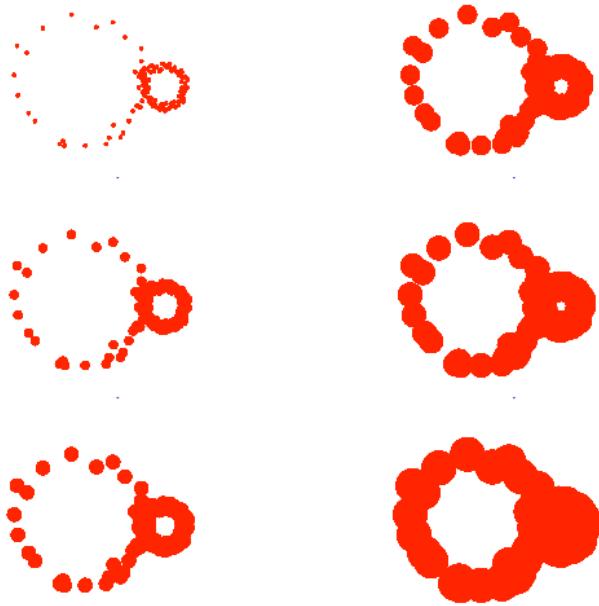


Figure 2.2: Example of constructing a filtration.

2.2.4 Persistent Homology

How to extend homology to finite metric spaces?

From the point cloud, a nested family of simplicial complexes, or a filtration, is constructed, parameterized by a filtration value ϵ , which controls the simplices present in the complex. The two most common ways of constructing a simplicial complex at each ϵ are the Čechcomplex and the Vietoris-Rips complex. 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.

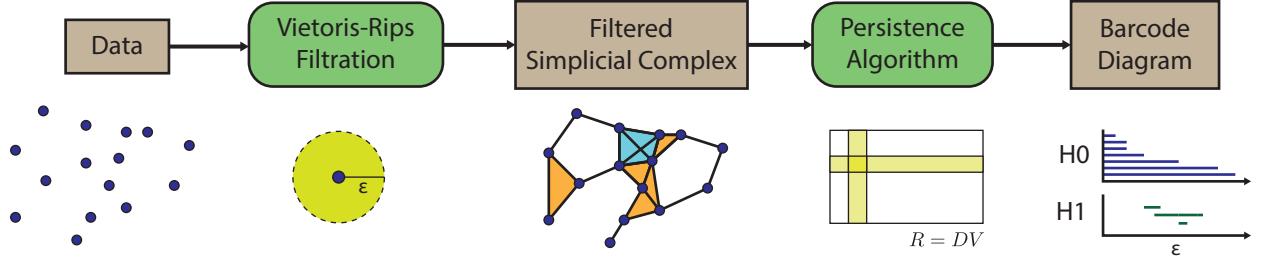


Figure 2.3: The Persistence Pipeline.

2.2.4.1 The Persistence Algorithm

While we act mostly as an end-user of persistent homology in this thesis, the algorithmics behind efficient computations of homology are interesting and worth including for comprehensiveness. Computing persistent homology is an exercise in linear algebra. The initial algorithms first induce a matching on a set of simplices. This is due to Zomorodian and is the implementation used in Javaplex and Dionysus. Then you reduce a couple of matrices. You can read off each bar and its representative cycle from looking at the zeros on this one particular matrix. Smith normal form.

More advanced algorithms have been developed that compute simplicial collapses: recognizing that the size of the simplicial set is often the limiting factor here, they collapse simplices into simpler structures that will have identical homology. This uses Discrete Morse Theory and is the idea behind implementations such as Perseus.

Include only simple implementation for $Z2$.

Several packages for computing persistent homology have been developed [Dionysus, Javaplex, Gudi, phom] and TDA frontend for R. Persistent homology is computed using Dionysus [63].

2.2.4.2 Stability

An important aspect of persistent homology is stability. Stability refers to how the output of persistent homology will change when the original data is perturbed, for example due to

noise or sampling. Will the existing bars change? Will new homology classes be formed? We would like the output of persistent homology to be stable under these perturbations. In general, our question is if I have some perturbation that takes my data from $D \rightarrow D'$, what can I say about the subsequent change in barcodes $B \rightarrow B'$? In general, if I have data D that is perturbed to new data D' , how will change Luckily, there is a result that bounds changes in the diagram, due to Chazal and coauthors [19]. After a few definitions, we state the stability theorem. First, we consider metrics on spaces.

Definition 1. The *Hausdorff distance* measures the distance between two shapes.

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\} \quad (2.5)$$

And $d_H(X, Y) = 0$ iff $X = Y$.

Definition 2. The *Gromov-Hausdorff distance* measures how far 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.

$$d_{GH}(X, Y) = \inf_{f,s} d_H(f(X), s(Y)) \quad (2.6)$$

Next, we consider how to define the distance between two persistence diagrams. To do so, we first need the concept of a *matching*. For two persistence diagrams A and B , a matching is a mapping from intervals in A to intervals in B , where we allow points to match to the diagonal to account for cases with unequal number of points. For each matched pair of intervals (a, b) , we define the L_∞ distance as

$$d_\infty(a, b) = \max\{|a_x - b_x|, |a_y - b_y|\}. \quad (2.7)$$

Definition 3. The *bottleneck cost* of a matching between two diagrams is the maximum L_∞ for all matched points. The *bottleneck distance* is defined to be the minimal bottleneck cost across all matchings. The matching with minimal bottleneck cost is the *bottleneck matching*.

$$d_B(A, B) = \inf_{n:A \rightarrow B} \sup_{x \in X} \|x - n(x)\|_\infty \quad (2.8)$$

The result of Chazal, Cohen-Steiner, Guibas, *et al.* [19] 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 .

Theorem 4. *The stability theorem.*

$$d_B(H_K(X), H_K(Y)) \leq d_{GH}(X, Y) \quad (2.9)$$

This bound establishes that small perturbations in the data will produce only small changes in the persistence diagram.

2.2.4.3 Statistical Persistent Homology

In persistent homology, the intuition is developed that long intervals are to be interpreted as large-scale, or robust, geometric features in data, while short intervals are more likely to correspond to noise or incomplete sampling.

More broadly, how can the information in the persistent diagram be used? Can this statement be made more precise? How short is short, and how will noisy sampling effect the observed diagram? When can a long interval be interpreted as a real feature, and can we assign measures of confidence to our estimates? 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.

Fasy and coauthors have developed ways of generating confidence intervals for persistence diagrams [35]. Based on some information about density, they can put a line off the diagonal below which points are to be considered noise (see example). Bubenik has developed the language of persistence landscapes [13, 12]. Several authors have examined the space of persistence diagrams as a Polish space, with notions of mean and variance. XXX et al have used the bootstrap to get estimates of the diagram robustness. Mukherjee et al.

Also see the work of Turner [84] and Mileyko [62].

- Probability measures on the space of persistence diagrams
- Functional summaries of the persistence diagram
- Confidence intervals on the persistence diagram
- Statistical inference using persistence diagrams
-

Wasserstein distance between diagrams.

$$W_p(D_1, D_2) = \left(\inf_{\gamma} \sum_{x \in D_1} \|x - \gamma(x)\|_{\infty}^p \right)^{1/p} \quad (2.10)$$

[Expand substantially, this section needs to be very strong.]

2.2.4.4 Multidimensional Persistence

First laid out in [16]. More work in [56]. Filtrations along different dimensions; how to relate?. Prototypical example: density and distance.

Our case is going to be slightly different. We will consider a set of points annotated with different metrics that we can put on it which will induce different homologies. Then we will see what happens we interpolate between those different metrics. [Discuss with Michael.]

2.3 Applying TDA to Molecular Sequence Data

Aligned genomic sequences can be naturally viewed as points in a high-dimensional sequence space. As more genomes continue to be sequenced, this space becomes more densely sampled. Using the standard genetic distance metrics described above, we can compute a pairwise distance matrix between genomes. This defines our finite metric space. From there, methods from TDA can be applied. In particular, we can apply persistent homology, and phylogenetic information contained in the dataset can be read off the resulting barcode diagram.

An important foundational point was described in [18]. In that paper, it was shown that if the evolutionary history contained in a particular data-set is tree-like, then there will be no higher homology in the resulting barcode diagram. In other words, the only nontrivial topology will appear in dimension zero.

If the set of genomes permits a phylogenetic representation, then positive-dimensional Betti numbers should vanish, since the topology of a tree is contractible. [[More about Gunnar's proof?](#)] 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.

We illustrate a simple example of how TDA can capture horizontal evolution from population data in Figure 2.4. Consider the reticulate phylogeny (Figure 2.4A): 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 2.4B, 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 2.4C we demonstrate the construction of a filtered simplicial complex, showing how the connectivity changes as the scale parameter ϵ is increased. Finally, in Figure 2.4D we see the resulting barcode diagram. Using H_0 we can 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 in a reticulation. In this way, we can analyze both the scale and frequency of reticulation in genomic data sets.

We summarize the connection between genomics and TDA in Table 2.1.

Table 2.1: 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

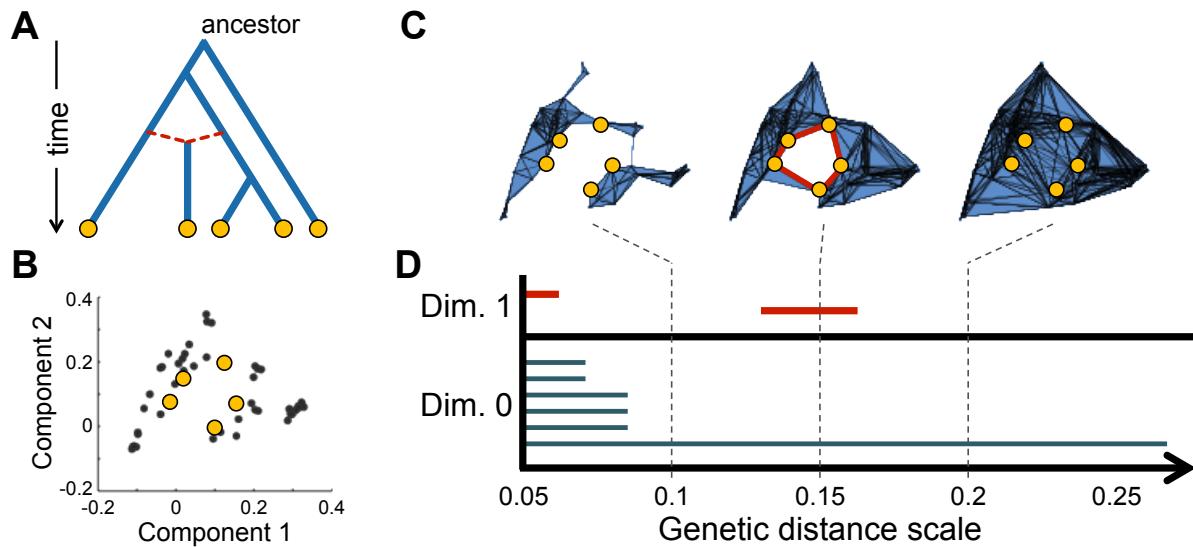


Figure 2.4: Applying persistent homology to genomic data. (A) An evolutionary genealogy including reticulation. (B) Data projected into 2-dimensions. (C) Construction of a filtered simplicial complex. (D) The resulting multiscale barcode diagram.

2.3.1 The Simplest Example

Recall the four gamete test, which tests for the presence of reticulate evolution by searching for the simultaneous presence of Recall the four gamete test. The simplest example is 00, 10, 01, 11. Under a Hamming metric this forms a loop.

2.3.2 The Space of Trees, Revisited

Recall our description of the space of trees in Section XX. There, we made reference to the idea that the goal of phylogenetic reconstruction can be seen as finding the best projection onto tree space for some arbitrary dataset. Expand our search space. New goal is to characterize the space of finite metric spaces. As we move off tree space, we know there is no additive tree that represents the data. Then we want some topological measure of deviation from additivity. So our updated picture looks like Figure XX.

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 [18], 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 as a 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 the 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 [7]. The approach is closely related to split decomposition, and it can be shown that the objects resulting from the two methods are identical [6]. The median graph approach imputes putative evolutionary ancestors into the set of vertices, and forms a network representing the incompatible 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 and 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 [42].

Additionally, we sought to more clearly interpret nontrivial higher homology (H_n for $n > 1$) in the barcode diagram. In Chan, Carlsson, and Rabadan [18], 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 interpretation 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 [18] 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

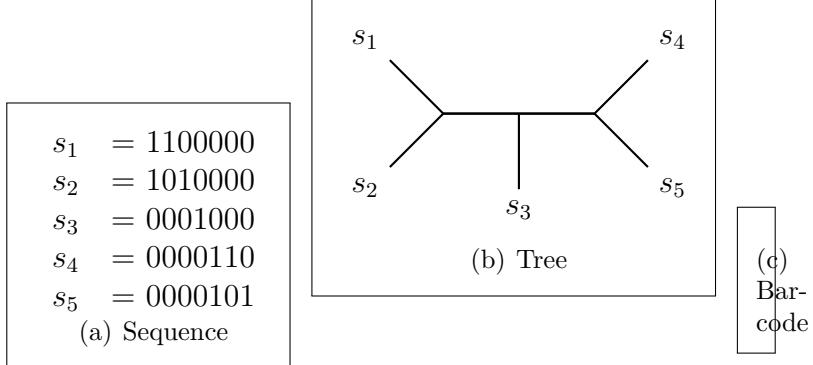


Figure 3.1: A tree is trivially contractible and has vanishing higher homology.

on the points such that

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

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 [18]. 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 transferred 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 [44].

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 [80]. 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 [80] 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 5. 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,

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

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 [31], but not using an explicit sequence representation. The filtration defined in Dress, Huber, and Moulton [31] is based on a complicated polytope construction scheme defined directly from the split decomposition. Given that all median graphs are split networks [45], 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

$$\text{Cl}(\mathcal{LC}) \hookrightarrow \text{Cl}(\mathcal{EC}) \hookrightarrow \text{Cl}(\mathcal{MC}) \quad (3.3)$$

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 [18] 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

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

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 [10].

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*.[39] 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

¹This 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

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

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 [42]. The homology of a cubical complex can be efficiently computed using the methods of Kaczynski, Mischaikow, and Mrozek [50] 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 Čech 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 [51]. The dataset consists of eleven sequences (nine unique) of the Adh locus from *Drosophila 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 [45].

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 [35, 9, 20]. 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

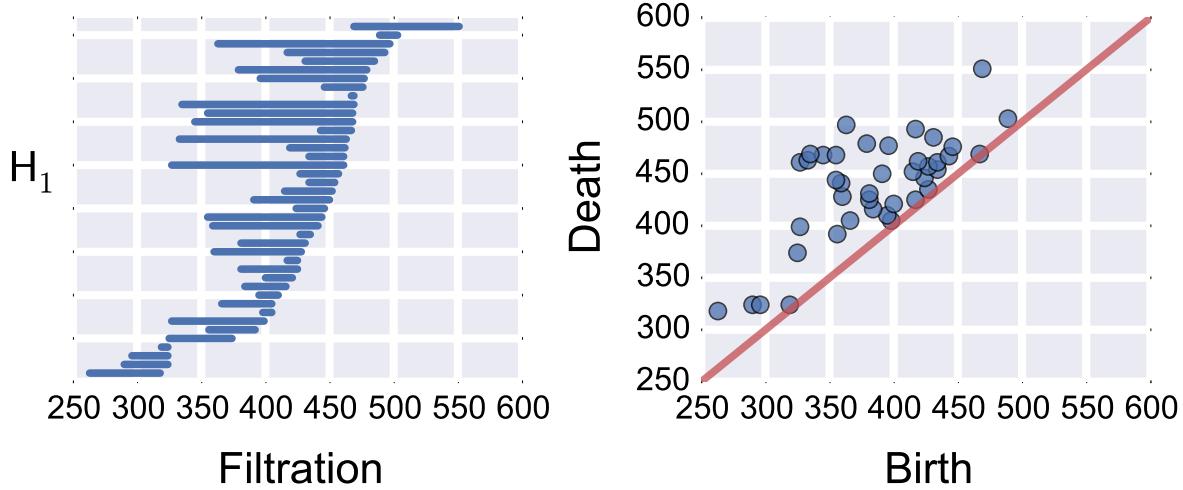


Figure 4.1: Two representations of the same topological invariants, computed using persistent homology. Left: Barcode diagram. Right: Persistence diagram. Data was generated from a coalescent simulation with $n = 100$, $\rho = 72$, and $\theta = 500$.

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 [85]. 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:

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

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 contractible tree structure, and is the component which we would like to quantify. Coalescent simulations were performed using `ms` [43].

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, \dots, b_K) ; the set of death times, (d_1, \dots, d_K) ; and the set of persistence lengths, (l_1, \dots, 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,

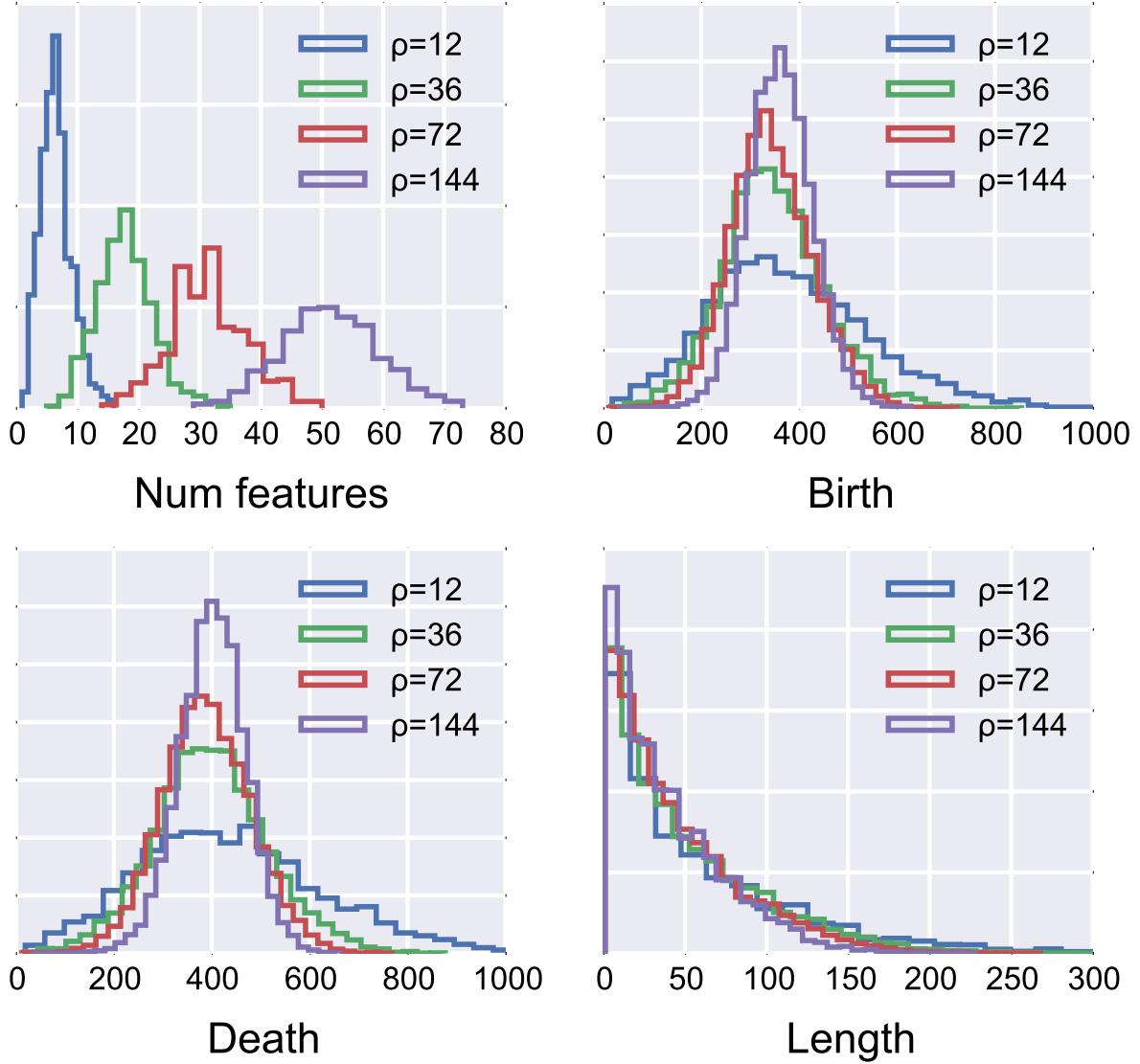


Figure 4.2: Distributions of statistics defined on the H_1 persistence diagram for different model parameters. Top left: Number of features. Top right: Birth time distribution. Bottom left: Death time distribution. Bottom right: Feature length distribution. Data generated from 1000 coalescent simulations 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

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

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

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

Birth times are Gamma distributed with shape parameter

$$\alpha = b_0 \rho + b_1 \quad (4.4)$$

and scale parameter

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

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 .

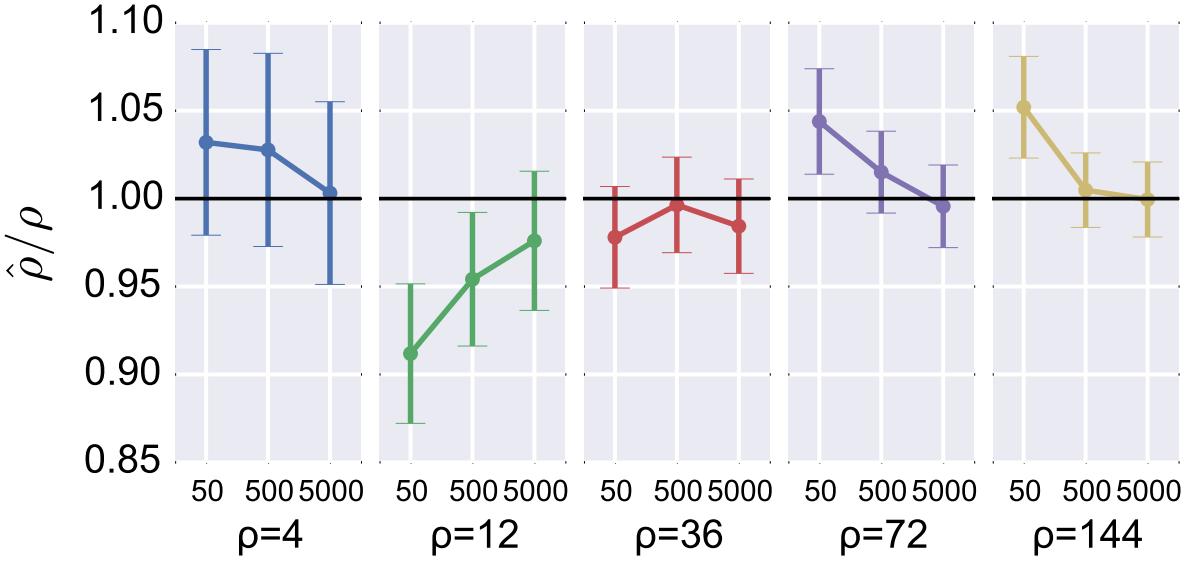


Figure 4.3: Inference of recombination rate ρ using topological information. The recombination rate ρ is estimated for five values $\{4, 12, 36, 72, 144\}$ at three different mutation rates $\{50, 500, 5000\}$. Mean estimate over 500 simulations and 95% confidence interval is shown.

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: Viruses and Bacteria

Chapter 5

Phage Mosaicism

5.1 Introduction

Phage: microbial viruses, infecting either bacteria [bacteriophages], archaea, or single celled eukaryotes. Bacteriophages, viruses whose host range are prokaryotic, are the most abundant organism on the planet. It is estimated that there are 10^{31} bacteriophages [75].

Because of the important role in the ecosystem by bacteria, it is especially important to understand the role of the viruses that infect them.

Inert intercellular transport form.

No single conserved gene that can be used to construct a taxonomy. Alternative approaches? Use full genome comparisons.

Bacteriophages, bacteria-infecting viruses, are the most abundant organism on the planet: 10^{31} organisms [cite]. [More background info: metagenomics, ocean phage, etc.]

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 [82]. Unlike the constituents of the generally accepted tree of life, phages lack a universal replicative machinery, rRNA, that can be used to define a species tree.

Table 5.1: Phage families defined by the ICTV

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

Nucleic acid composition is either double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), or single-stranded RNA (ssRNA). Of these ds-DNA is by far the most common.

Morphological classification is primarily based on head/capsid shape and tail length.

Table 5.1 presents an overview of phage families defined by the ICTV.

It has long been known that phage species are genetic mosaics with extremely high rates of lateral exchange. The advent of genome sequencing solidified this observation and brought to bear questions about the applicability and interpretation of the ICTV taxonomy. Unlike prokaryotes and eukaryotes, phages do not have ribosomal RNA. Indeed there is substantial debate over whether phages are truly alive and whether or not they should be a component in the tree of life. Given the substantial amount of genetic diversity

The current bacteriophage taxonomy is inconsistent with recently collected genomic data. In Figure 5.1 we see three different bacteriophage species. 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

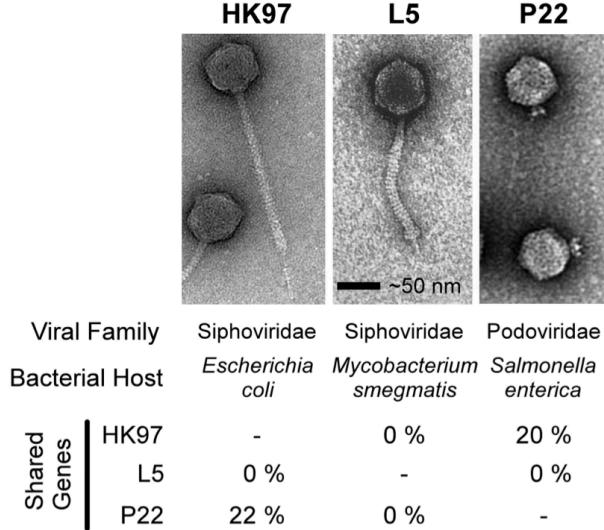


Figure 5.1: Inconsistency of genomic and morphological approaches. HK97 and L5 are classified under same viral family, despite sharing no homologous genetic content. P22 on the other hand shares many genes. Figure adapted from [54]

that HK97 and L5 should indeed be classified as distinct from P22. However, genomic analysis indicates that HK97 and L5 share no gene content. Despite belonging to different viral families, HK97 and P22 share 20% gene content. If we are to take genomic data as the core information defining.

Alternatives have been proposed based on whole genome analysis. For example, see Rohwer and Edwards and the phage proteomic tree [74]. However, these models still broadly assume a tree like structure.

In this chapter, we use topological approaches to define a systematic way of structuring phage relationships based on gene content. This work is based on data collected by Lima-Mendea *et al.* [58] and Kristensen *et al.* [52]. First, we use persistent homology to characterize reticulation in phage genomes. We find H_0 consistent with existing phage taxonomies. We interpret H_1 as evidence for genetic exchange due to shared ecology and host-range. Second, we visualize phage relationships using mapper, identifying non-obvious relationships between phages of varying nucleic acid content.

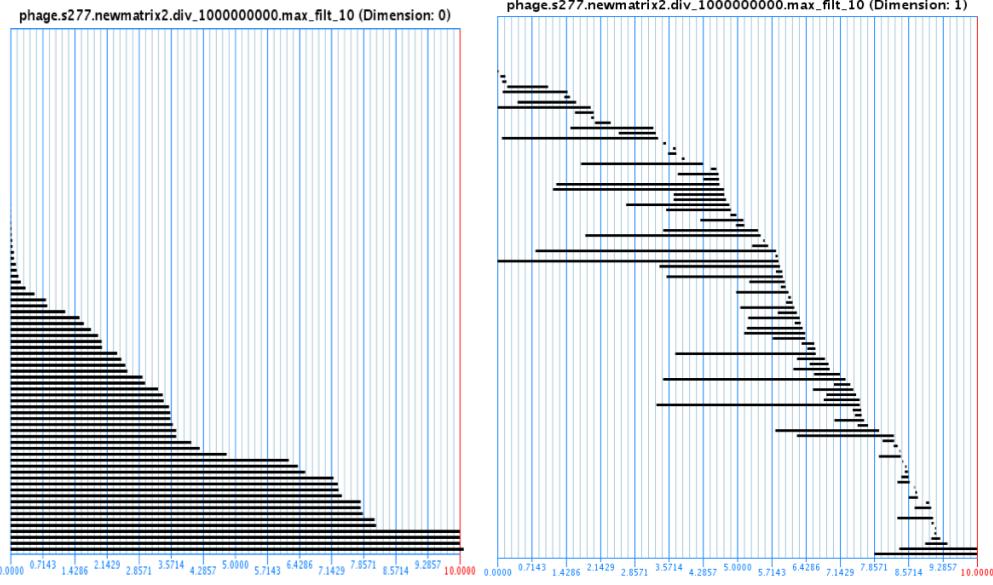


Figure 5.2: Bacteriophage Barcode Diagram using the Lima-Mendez dataset.

5.2 Approach

5.2.1 Data

First data set follows that from [58] Input data is 306 bacteriophage genomes. 250 dsDNA, 36 ssDNA, 12 dsRNA, and 8 ssRNA. 1,9537 genes clustered into 8,576 gene families using BlastP. Construct Phyletic profile: $n \times p$ binary gene presence/absence matrix. 29 outlier phages, discard and keep only subset S277.

Second data set follows that from [52] and is more extensive. This dataset defines phage orthologous groups (POGs) and includes viruses of prokaryotic host including bacteria and archaea. Input data is 1,005 phage genomes.

5.3 Results

We show the barcode diagrams in Figure 5.2. H_0 gives an initial classification. Cycles in H_1 can be mapped to specific reticulation patterns. These cycles reflect shared environment.

In Figure XXX we show the barcode diagram. We show the relationship of these bars

within an existing phage taxonomy. Extract multiscale patterns of reticulation. Phage phylogeny taken from [40].

Finally, we use Ayasdi Iris to construct a network visualization of the phage phyletic profiles. We see several interesting things. Using Ayasdi, we can identify gene enrichment in particular clusters. We can also correlate with lifestyle and ecological properties.

5.4 Phage Ecological Properties

We can associate particular generators with ecological groupings.

5.5 Conclusions

In this chapter we have analyzed bacteriophage mosaicism.

Chapter 6

Reassortment in Influenza Evolution

6.1 Introduction

In this chapter we analyze influenza A virus, a common human pathogen. Influenza virus is a segmented single-stranded RNA virus of family Orthomyxoviridae.

It is well-known that frequent reassortment of viral segments punctuates the evolution of influenza [cites].

Since the influenza pandemic of 2009, substantial effort has been put into. The NCBI Influenza Virus Resource now contains over 400,000 unique viral isolates. The large quantity of genomic data that has been collected provides an ideal test bed for studying evolution in the reticulate limit.

6.2 Influenza Virus

In this chapter we analyze influenza A virus, a common human pathogen. Influenza is a segmented single-stranded RNA orthomyxovirus. It is well known that frequent reassortment punctuates the evolution of influenza. Reassortant strains can lead to antigenic novelty when ingernal segments adapted to humans pair up with novel external segments. Such events led to the human influenza pandemics of 1957 and 1968. Influenza provides a useful testbed for

detection of reticulate evolution because of the large quantity of data that has been collected.

Influenza is a single-stranded RNA virus of family Orthomyxoviridae. Influenza has a segmented genome with eight segments coding for 11 proteins. The segments are generally ordered from longest to shortest.

HA is the surface antigen that is responsible for binding the viral particle to the cell. HA binds to sialic acid groups on the cell surface. 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. Human and avian adaptations. NA is the surface protein that cleaves the replicating virus from the cell surface. Together, HA and NA determine the subtype of virus and are a first marker of host transmission and specificity.

PB1, PB2, and PA form the 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.

What is the estimated mutation rate of influenza?

Natural host reservoir: aquatic birds.

Origin of 2009 H1N1 pandemic: triple reassortment. Origin of 2013 H7N9 pandemic: triple reassortment. Also discuss H5N1 and H5N2.

Therefore, ways of automatically characterizing and representing reassortment are necessary.

We applied methods from TDA to data from more than 3,999 avian influenza genomes from the NIH Influenza Sequence Database.

We applied persistent homology to data for more than 3,000 avian influenza genomes from the NIH Influenza Sequence Database. Examining each genome segment separately, we recovered only zero-dimensional homology, consistent with no intra-segmental recom-

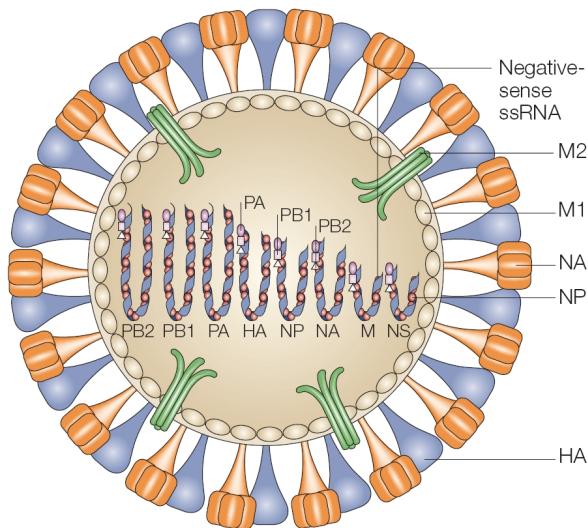


Figure 6.1: Structure of an influenza virus particle. The surface capsid is formed from matrix proteins M1 and M2. Surface antigens HA and NA coat this surface and are involved in cell entry and exit. PB1, PB2, and PA form a polymerase complex assisting in

bination. In other words, segment-by-segment the evolution of influenza is tree-like and amenable to a phylogenetic tree representation.

However, a similar analysis of the concatenated full genome reveals a complex topology, with a large number of loops in one and two dimensions.

6.3 Reassortment

The evolution of influenza is punctuated by frequent reassortment. To characterize influenza evolution, we computed the persistent homology of four influenza datasets from avian, swine, and human hosts, each numbering as many as 1,000 genomic sequences. When applied to a single segment of the virus unaffected by reassortment, higher-dimensional homology groups vanish (Fig. 2). Alignments of single segments are therefore suitable for phylogenetic analysis. 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.

(More explanation of this figure!!)

6.4 Multiscale Flu Reassortment

To test our model on biological data, we considered reassortment in avian influenza virus. Influenza is a single-stranded RNA virus that is naturally found in avian populations. Each viral genome has eight genetic segments. Subtypes are defined by two segments, hemagglutinin (HA) and neuraminidase (NA), e.g. H1N1 and H3N2. When a host cell is coinfecte

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 6.3, along with density estimates for the birth and death distributions. Both birth and death

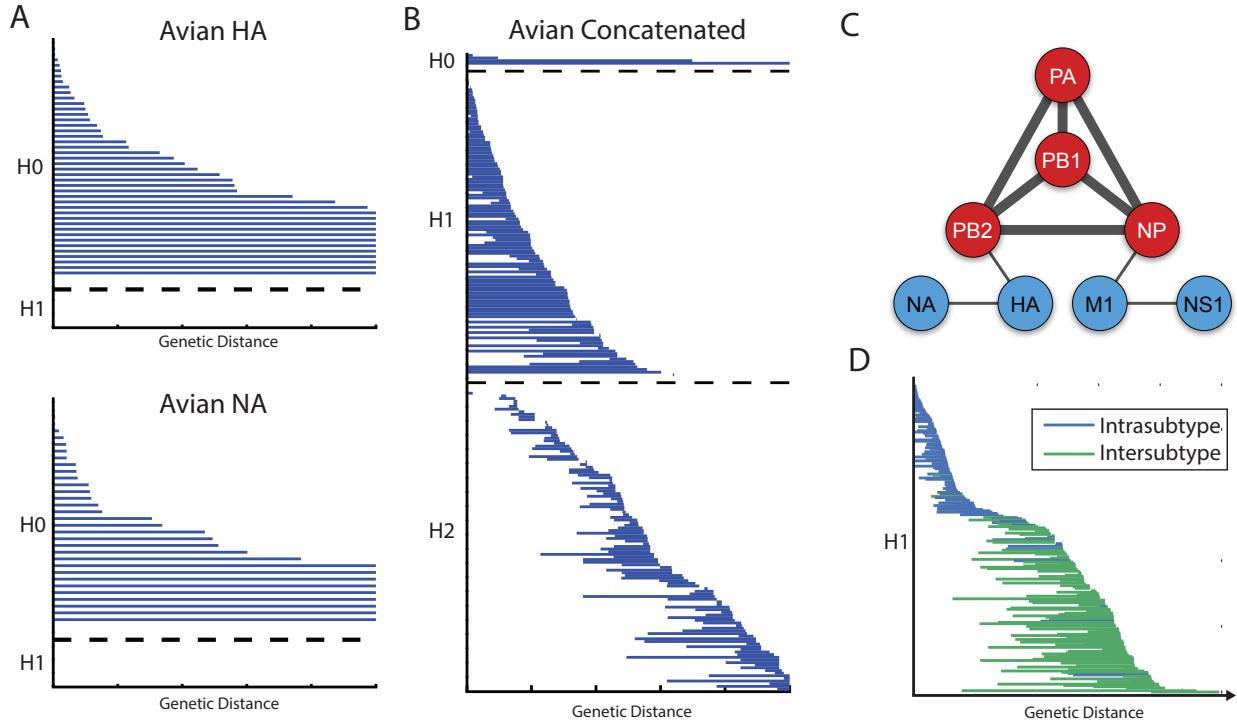


Figure 6.2: flu..

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 6.3 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 occurring 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

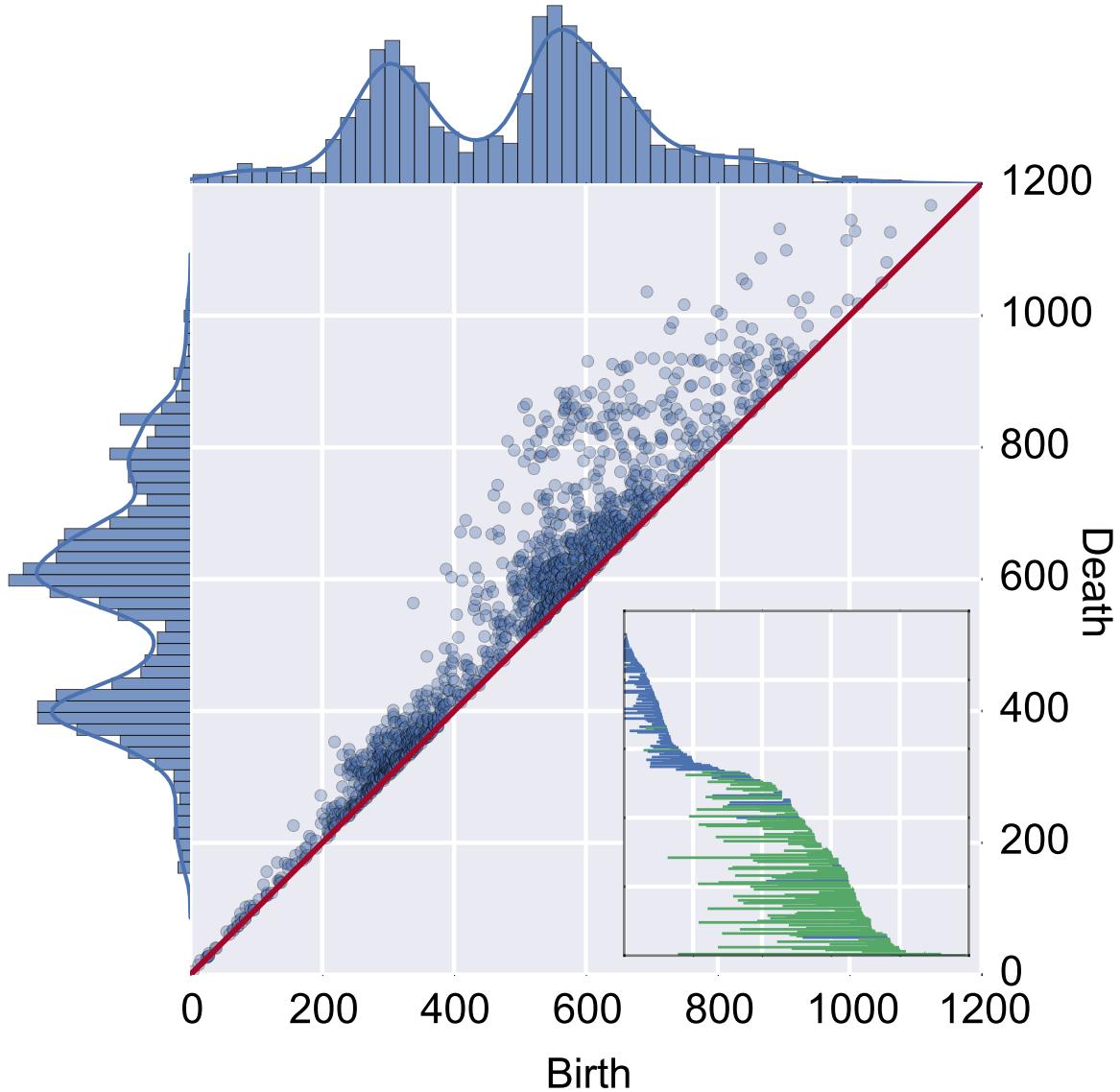


Figure 6.3: The H_1 persistence diagram computed from an avian influenza dataset. On the top and left are plotted the marginal distributions of birth and death times, along with a density estimate for each distribution. The bimodality indicates two scales of topological structure. Inset: The barcode diagram for a subset of this data. Blue bars have representative cycles involving only one subtype, green bars have cycles involving multiple subtypes.

further study. This illustrates a real world example in which multiscale topological structure can be captured by persistent homology and given biological interpretation.

6.5 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.

6.6 Conclusions

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 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 estimated reassortment rates. We have estimated recombination rates. Further, from the persistence diagram we identified a bimodal presence of H1 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 presents 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 [66, 83, 70]. 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. Three mechanisms of horizontal transfer are

identified, depending on the route by which foreign DNA is acquired [68]. 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 [73]. 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.

The presence of horizontal gene transfer precludes accurate phylogenetic characterization, because different segments of the genome will have different evolutionary histories. Bacterial species definitions and taxonomic classifications are made on the basis of 16S ribosomal RNA, a highly conserved genomic region between bacteria and archaea species [89]. However, the region generally accounts for less than 1% of the complete genome, implying that the vast majority of evolutionary relationships are not accounted for in the taxonomy [24]. Because of the important role played by lateral gene transfer, new ways of characterizing evolutionary and phenotypic relationships between microorganisms are needed.

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 [37]. 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 [70]. Using persistent homology, we characterize the rate and scale of recombination in the core genome using multilocus sequence data. To

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

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

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 [78]. We identify two major populations of *S. aureus*, and observe one cluster with strong enrichment for the antibiotic resistance gene *mecA*. Importantly, 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 resistance 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 individual has a unique microbiome, and we speculate that microbiome typing of this sort may be 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 exchange 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 [49]. 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 $\nu = 0$ [26]. The computation is performed using javaplex [81]. 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

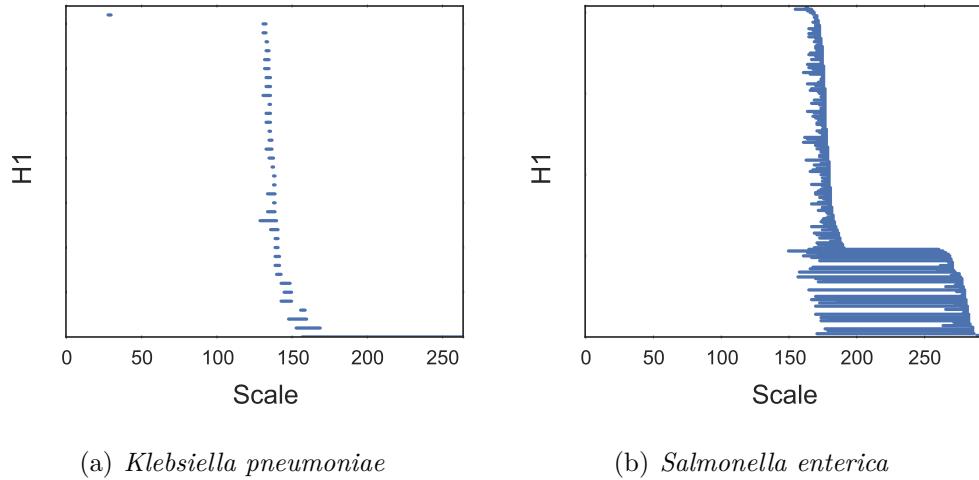


Figure 7.1: Barcode diagrams reflect different scales of core genomic exchange in *K. pneumoniae* and *S. enterica*.

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. pylori* 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 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.

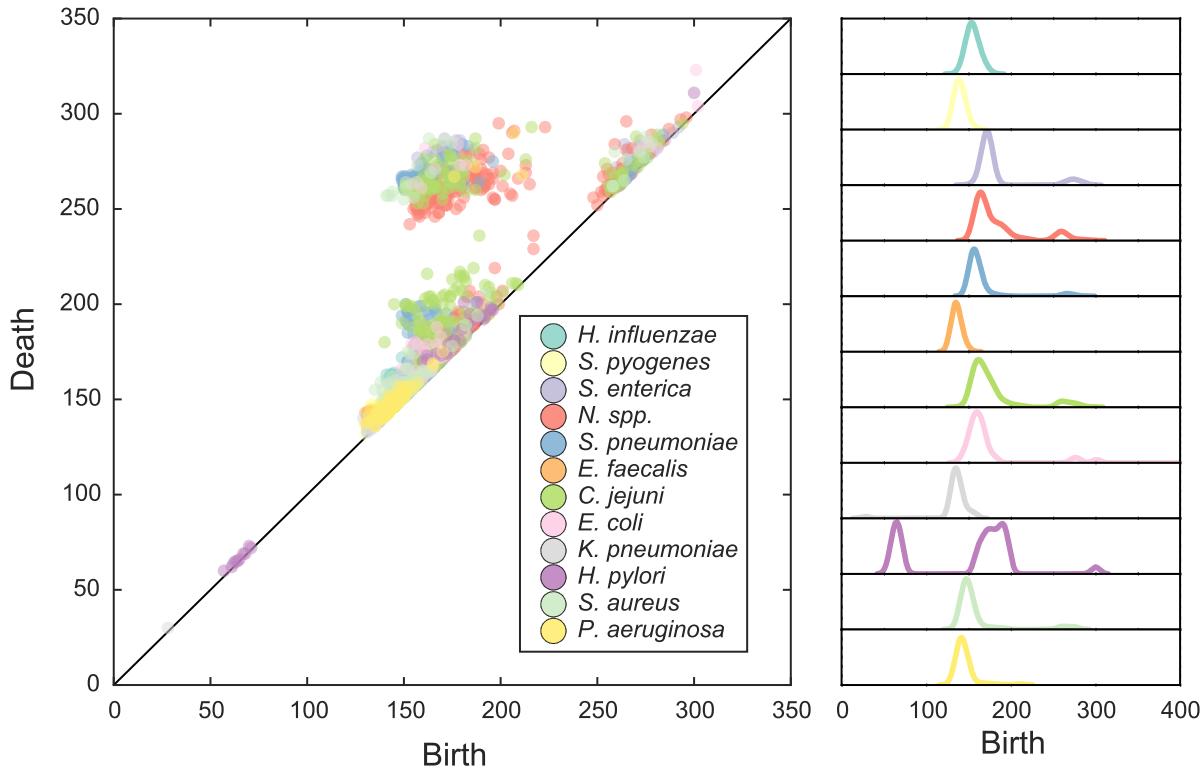


Figure 7.2: The H_1 persistence diagram for the twelve pathogenic strains selected for this study using MLST profile data. There are three broad scales of recombination. To the right is the birth time distribution for each strain. *H. pylori* has an earlier scale of recombination not present in the other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

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 annota-

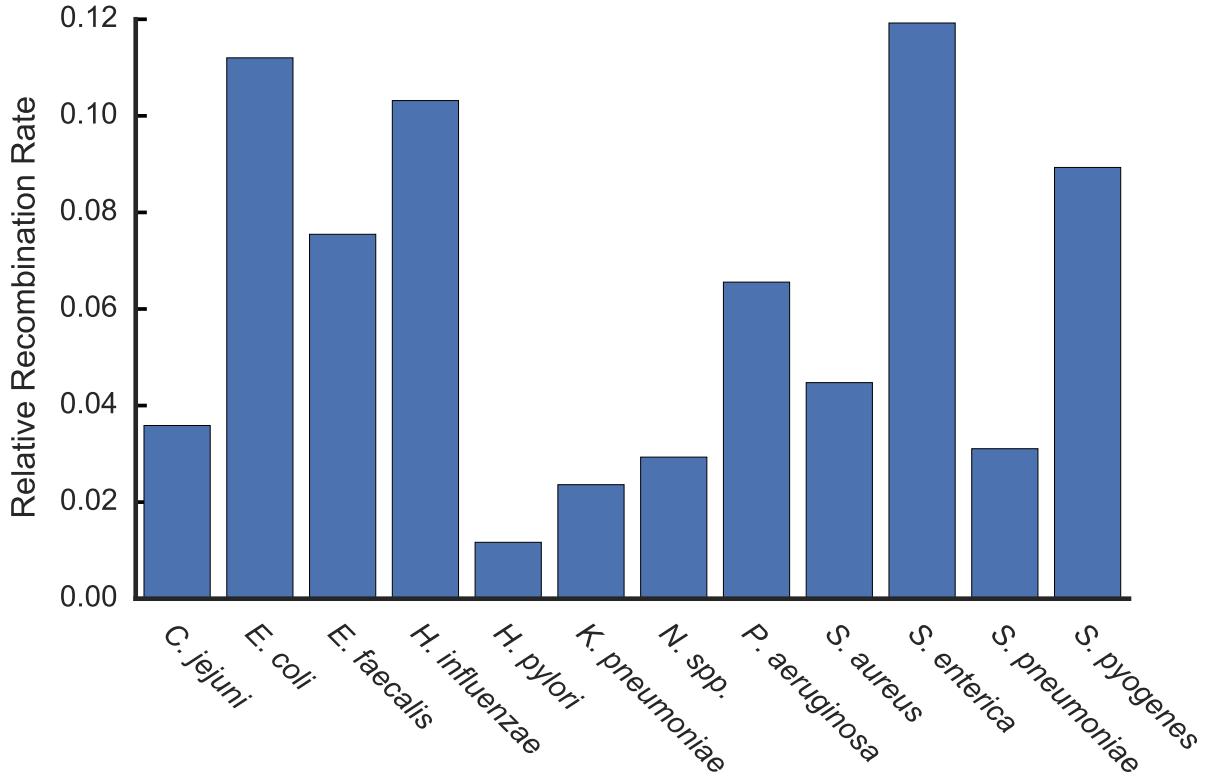


Figure 7.3: Relative recombination rates computed by persistent homology from MLST profile data.

tions [88]. The FigFam annotation scheme consists of over 100,000 protein families curated from over 950,000 unique proteins [61].

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 [37]. 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

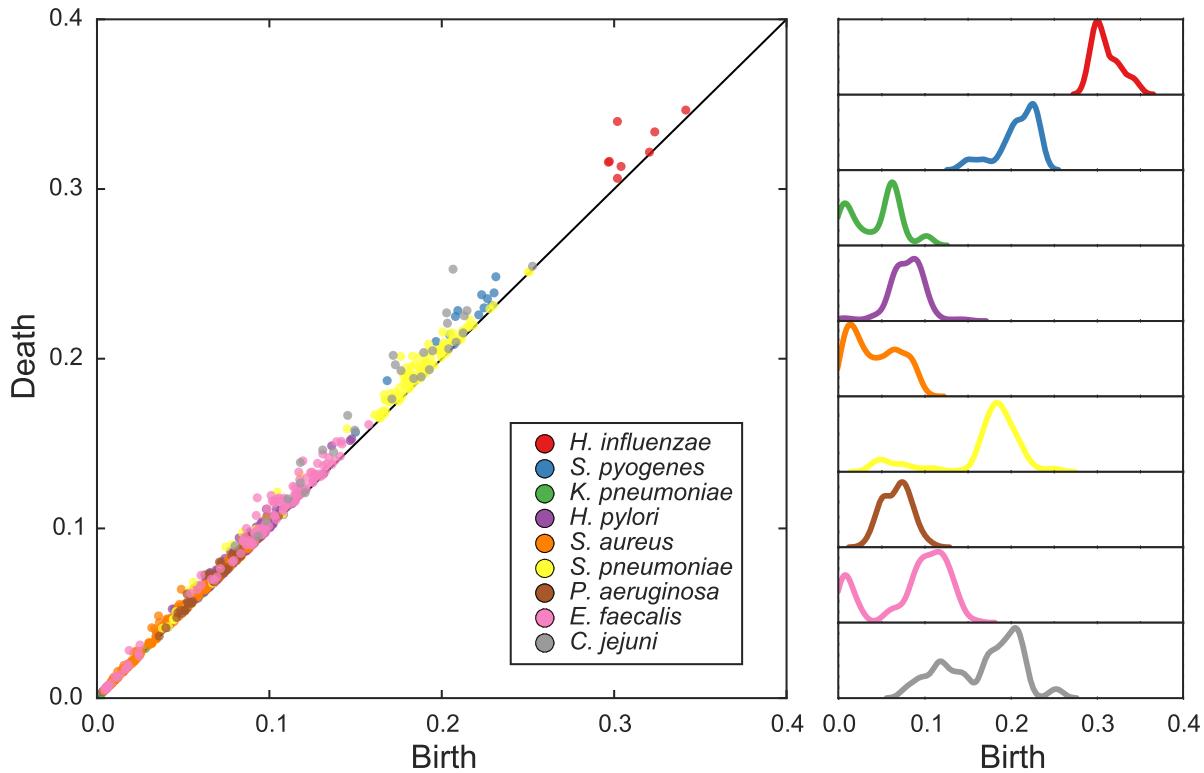


Figure 7.4: Persistence diagram for a subset of pathogenic bacteria, computed using the FigFam annotations compiled in PATRIC. Compared to the MLST persistence diagram, the Figfam diagram has a more diverse scale of topological structure.

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, particularly in the hospital setting. The emergence of antibiotic resistant *S. aureus* (MRSA) strains 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 Staphylococcal cassette chromosome *mec* (*SCCmec*). *mecA* codes for a

dysfunctional penicillin-binding protein 2a (PBP2a), which inhibits β -lactam antibiotic binding, the primary mechanism of action [47]. 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 [5]. 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 [23].

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

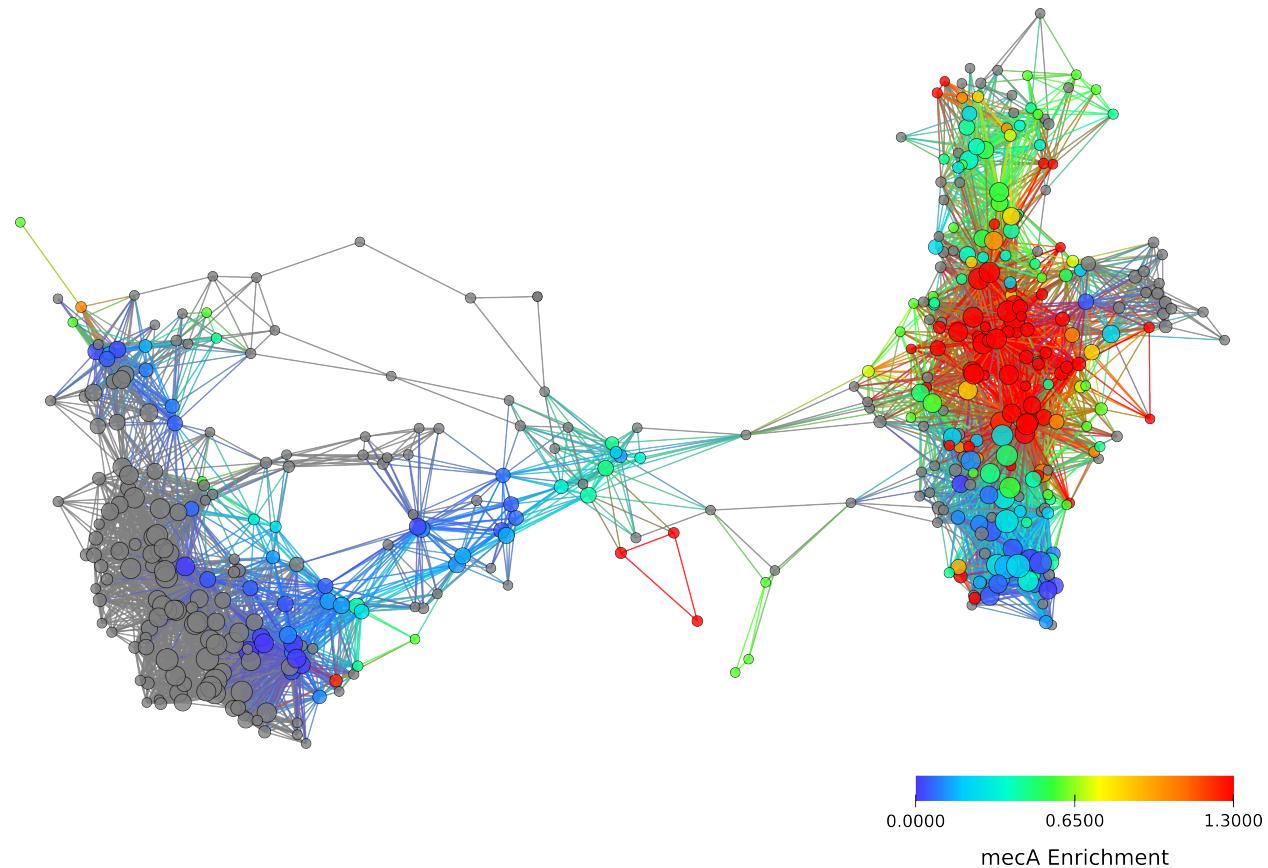


Figure 7.5: The FigFam similarity network of *S. aureus* constructed using Mapper as implemented in Ayasdi Iris. We use a Hamming metric and Primary and Secondary Metric SVD filters (res: 30, gain 4x, eq.). Node color is based on strain enrichment for *meca*, the gene conferring β -Lactam resistance. 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.

symbiotically within a human host, can act as a reservoir of antimicrobial resistance genes [79, 71]. 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 [22]. 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 be-

cause 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 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 phyla, 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 Proteobacteria is clinically 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 [22]. In this case, an individual's 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.

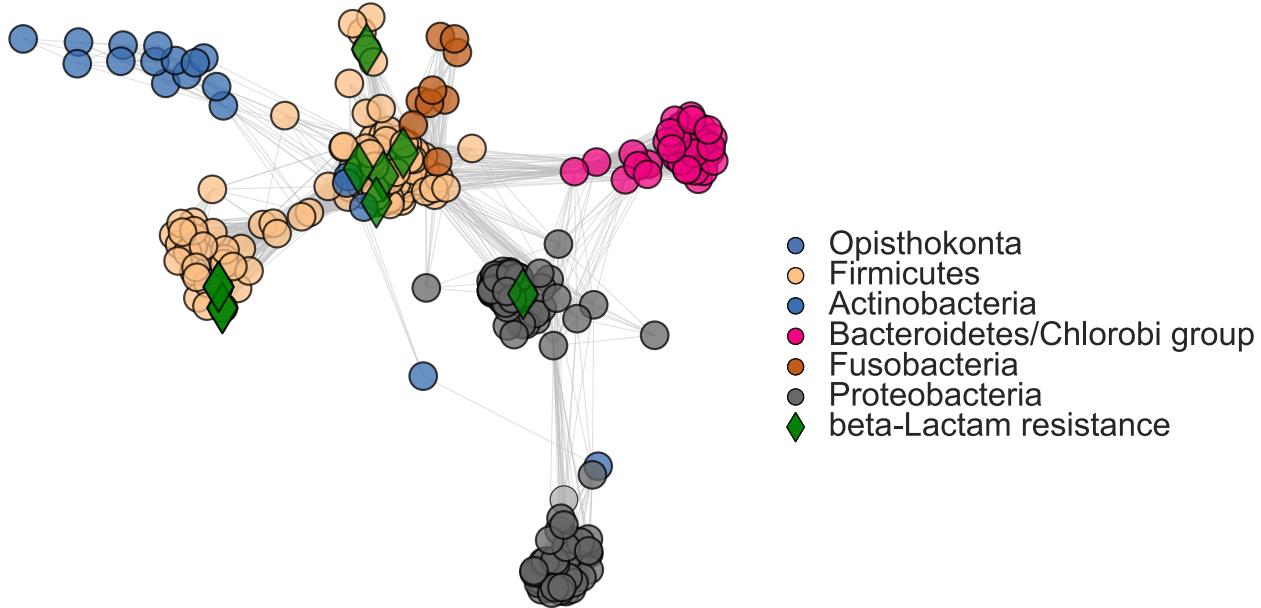


Figure 7.6: The FigFam similarity network of gastrointestinal tract reference strains identified in the Human Microbiome Project. The green diamond identifies the strains carrying resistance to β -Lactam antibiotics.

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

Prokaryote Reticulate Evolution - Tree of Life

In this chapter we examine evolutionary relationships across the prokaryotic domain. As input data, we use the Cluster of Orthologous Genes (COG) database at NCBI [38]. Using a combination of topological tools, we present a construction meant to extend the tree of life paradigm.

8.1 Introduction

In this chapter, we examine evolutionary relationships across the prokaryotic domain.

First, we use persistent homology to characterize reticulation. Second, we use mapper to visualize evolutionary relationships.

8.2 Materials and Methods

As input data, we use the Cluster of Orthologous Genes (COG) database from NCBI [38]

8.3 Results

To visualize relationships, we use the Mapper algorithm, as implemented in Ayasdi Iris.

8.4 Conclusion

In this chapter, we have examined evolutionary relationships across the prokaryotic domain.

Part III

Applications: Human Data

Chapter 9

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.

9.1 Introduction

9.2 Material

9.3 Results

9.4 Conclusions

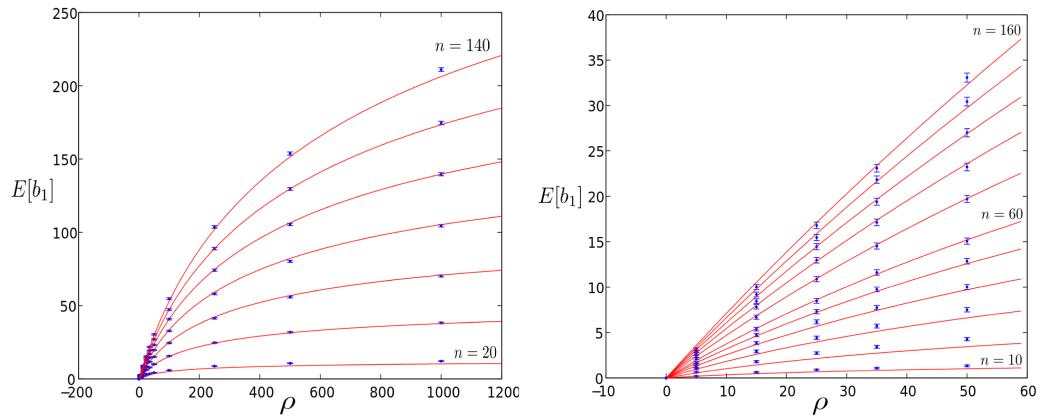
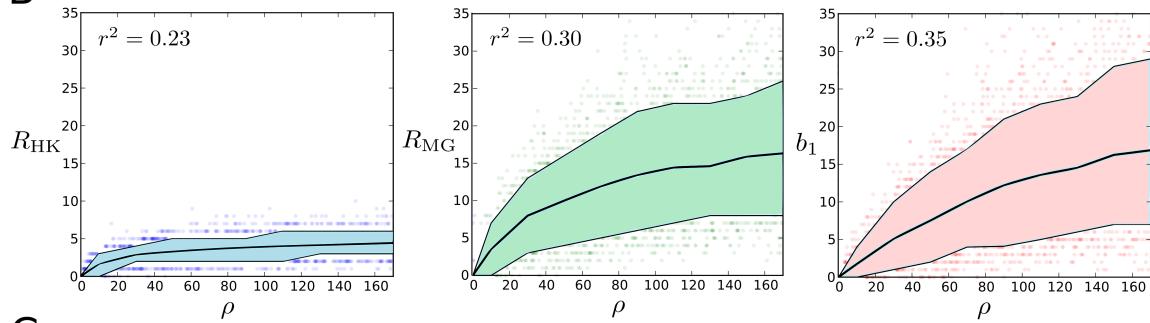
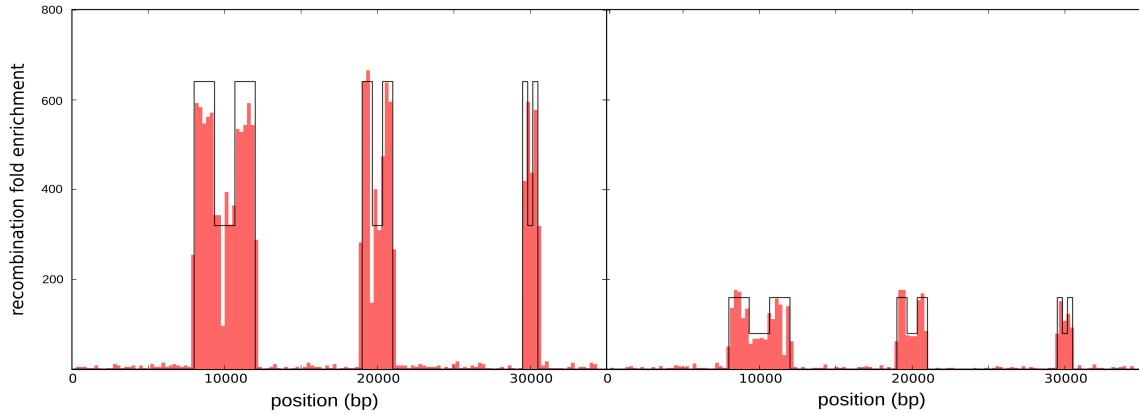
A**B****C**

Figure 9.1: Calibration

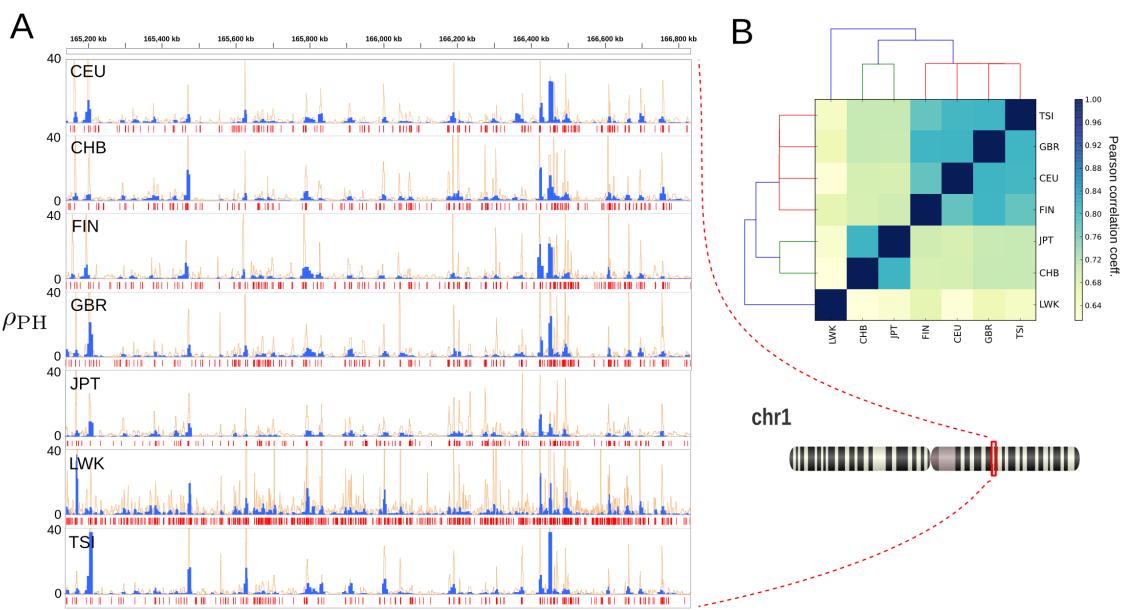


Figure 9.2: Population Tracks

Chapter 10

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.

10.1 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 [27]. Chromatin conformation is dynamic, and will change throughout the cellular cycle, under the influence 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 [57, 27, 4]. 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 [4]. 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 [18, 32]. 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.

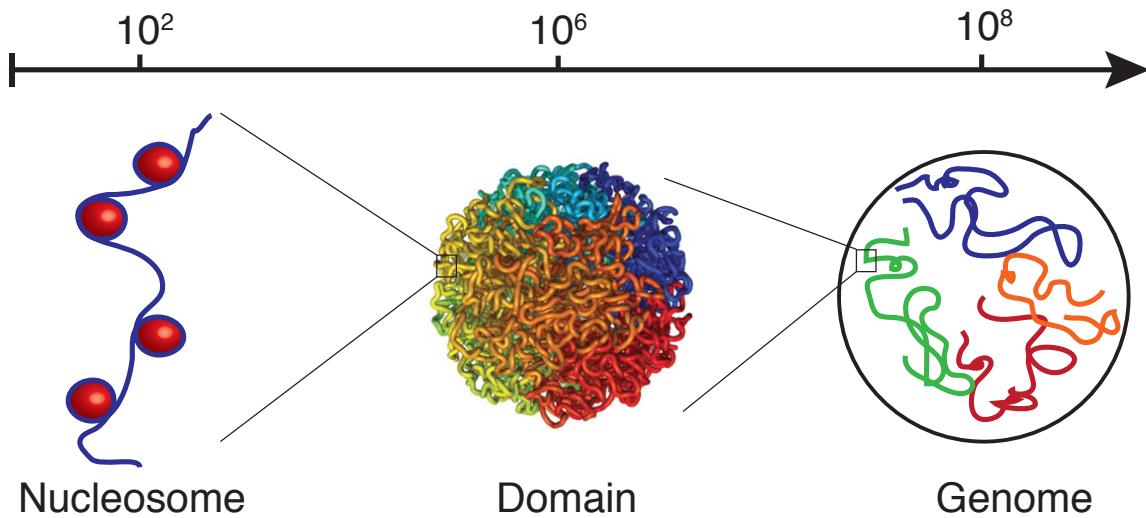


Figure 10.1: Three hierarchies of chromatin organization. At the 100 bp scale, DNA chains wrap around protein complexes called nucleosomes. At the megabase scale, these chains are compacted into domains. Lieberman-Aiden *et al.* proposed closed domains form a *fractal globule* structure. At the genome scale, chromosomes fold into the nucleus in separate territories. The fractal globule represents densely packed regions not open for transcription. Fractal globule image from [57]. Reprinted with permission from AAAS.

10.2 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 [27]. 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 [4] 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 [15] 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 [60], we define the *size* of a PH class as

$$x_i = \frac{b_i + d_i}{2}. \quad (10.1)$$

The distribution of PH class sizes reflects the scales of folding observed. We use Dionysus to compute persistent homology [63].

10.2.1 Long-Range Chromatin Interactions

Long-range chromatin interactions can manifest in a number of different biological consequences at megabase scales. In Figure 10.2 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 nonadjacent loci.

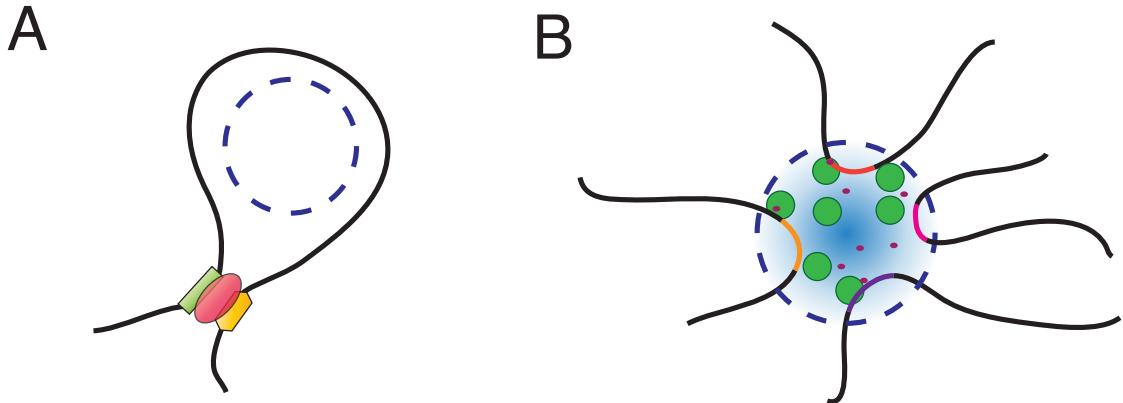


Figure 10.2: Two examples of long range chromatin interactions resulting in a topological loop. (A) A protein mediated (red) point-interaction between an enhancer (green) and promoter (yellow) sequence. (B) A transcription factory consists of dense RNA polymerase (green) around a structural core in which adjacent genomic loci (colored segments) will be cotranscribed. Transcription factors (purple) are shown.

10.2.2 Minimal Cycle Algorithm

It is important to be able to localize a cycle in order to annotate particular loops and 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 [77]. This does not uniquely specify a cycle, and we break ties by preferring cycles with shorter jumps. Specifically, if $x y \dots z$ is homologous to $x x + 1 y \dots z$ (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.

10.3 Polymer Simulations

To explore the use of topological methods for analyzing chromatin data, we used code from [30] 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 10.3 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.

10.4 Caulobacter Data

We examined interaction data from *Caulobacter crescentus* as published in [55]. *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 sample GSM1120446, a wildtype *Caulobacter* cell. In Figure 10.4A 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 10.4B is the barcode diagram computed from this contact map. Finally, in

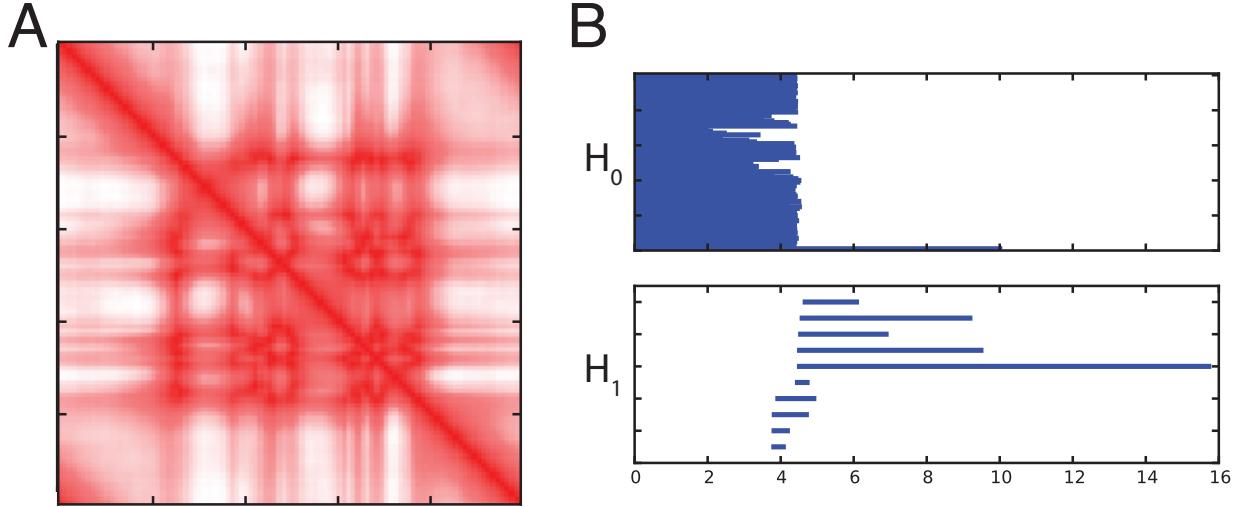


Figure 10.3: Polymer Simulation. (A) 50 Mb polymer with 10 fixed loops is allowed to reach an equilibrium conformation. (B) PH identifies 10 H_1 loops.

Figure 10.4C 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 10.5 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 10.4C. 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 [86]), 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 [55], but rather reflect larger scale folding patterns.

10.5 Human Data

We examined one of the original human Hi-C data sets as published in [57]. 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

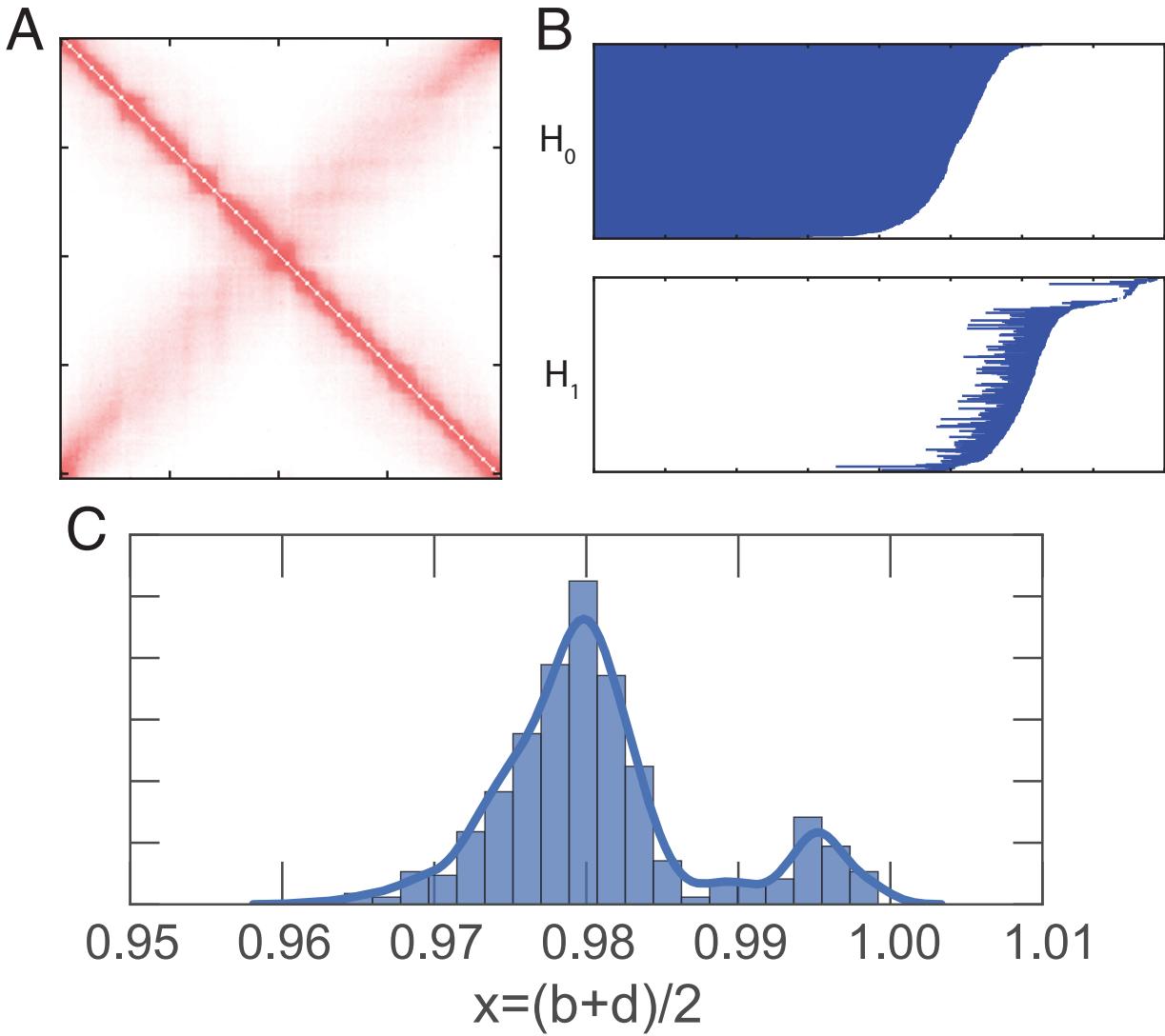


Figure 10.4: (A) Contact map and (B) barcode diagram for *Caulobacter*. (C) Distribution of H_1 bar sizes for *Caulobacter* shows a bimodal scale of folding patterns.

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 10.6 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 .

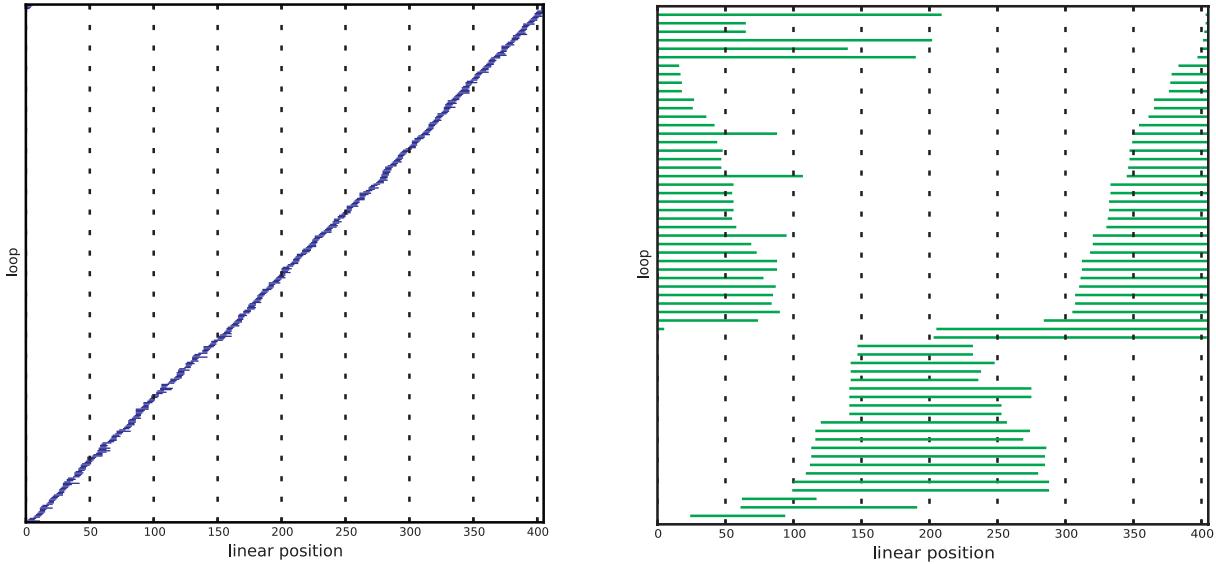


Figure 10.5: Minimal cycles projected linearly for *Caulobacter*. Left: small loops distribute uniformly across the genome. Right: pattern of large loops, which segregate into two chromosomal domains.

In Figure 10.7 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 [57], 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 [48, 72], 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.

10.6 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

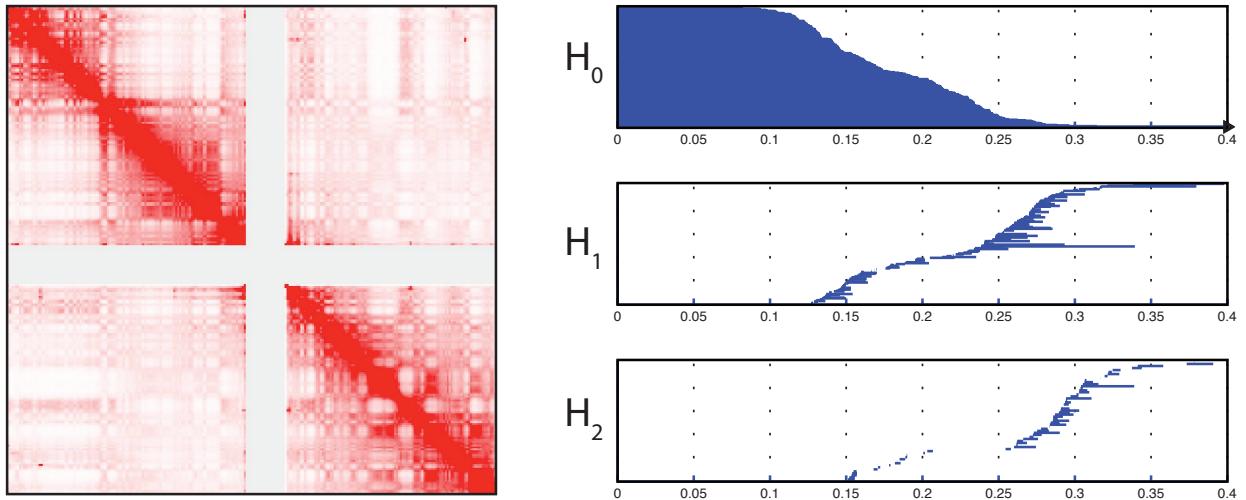


Figure 10.6: Hi-C data chromosome 1 from GM06690 human cell line data, from [57]. Left: Contact map representation. Right: PH identifies complex multiscale topology.

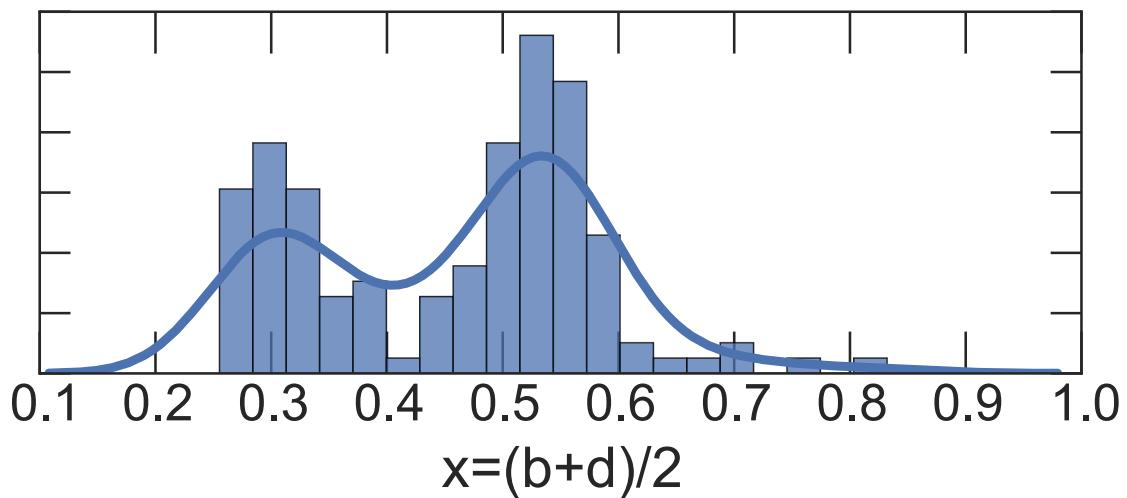


Figure 10.7: Distribution of H_1 bar sizes for GM06690 human cell line data shows a bimodal scale of folding patterns.

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. Interchromosomal 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 11

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](https://arxiv.org/abs/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.
- [3] M. L. Arnold, *Evolution through Genetic Exchange*. Oxford University Press, 2007.
- [4] 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](https://doi.org/10.1186/s13059-015-0745-7).
- [5] Ayasdi Inc., *Ayasdi core*, <http://www.ayasdi.com>, 2015.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] A. J. Blumberg, I. Gal, M. A. Mandell, and M. Pancio, “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.
- [10] K. Borsuk, “On the imbedding of systems of compacta in simplicial complexes,” *Fundamenta Mathematicae*, vol. 35, no. 1, pp. 217–234, 1948.
- [11] P. J. Bowler, *Evolution: The History of an Idea*. University of California Press, 2003.
- [12] P. Bubenik, “Statistical topological data analysis using persistence landscapes,” *Journal of Machine Learning Research*, vol. 16, pp. 77–102, 2015.

- [13] P. Bubenik and P. T. Kim, “A statistical approach to persistent homology,” *Homology, Homotopy and Applications*, vol. 9, no. 2, pp. 337–362, 2007.
- [14] D. Burke, “Recombination in hiv: an important viral evolutionary strategy,” *Emerging Infectious Diseases*, vol. 3, no. 3, pp. 253–259, Sep. 1997.
- [15] G. Carlsson, “Topological pattern recognition for point cloud data,” *Acta Numerica*, vol. 23, pp. 289–368, 2014.
- [16] G. Carlsson and A. Zomorodian, “The theory of multidimensional persistence,” *Discrete & Computational Geometry*, vol. 42, no. 1, pp. 71–93, 2009.
- [17] 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.
- [18] 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](https://doi.org/10.1073/pnas.1313480110).
- [19] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot, “Gromovhausdorff stable signatures for shapes using persistence,” in *Computer Graphics Forum*, Wiley Online Library, 2009, pp. 1393–1403.
- [20] 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.
- [21] 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.
- [22] T. H. M. P. Consortium, “Structure, function and diversity of the healthy human microbiome,” *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.
- [23] 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.
- [24] T. Dagan and W. Martin, “The tree of one percent,” *Genome Biology*, vol. 7, no. 10, p. 118, 2006.
- [25] C. Darwin, *On the Origin of Species by Means of Natural Selection*. London: Murray, 1859.

- [26] 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.
- [27] 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.
- [28] 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](https://doi.org/10.1126/science.284.5423.2124).
- [29] 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](https://doi.org/10.1186/gb-2006-7-9-116).
- [30] 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.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] 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](https://doi.org/10.1214/14-AOS1252).
- [36] W. M. Fitch and E. Margoliash, “Construction of phylogenetic trees,” *Science*, no. 3760, pp. 279–284, 1967. DOI: [10.1126/science.155.3760.279](https://doi.org/10.1126/science.155.3760.279).
- [37] 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](https://doi.org/10.1126/science.1127573).

- [38] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin, “Expanded microbial genome coverage and improved protein family annotation in the COG database,” *Nucleic Acids Research*, pp. D261–D269, 43 2014. doi: [10.1093/nar/gku1223](https://doi.org/10.1093/nar/gku1223).
- [39] B. Gärtner, “Fast and robust smallest enclosing balls,” in *Algorithms-ESA 99*, Springer, 1999, pp. 325–338.
- [40] G. Glazko, V. Makarenkov, J. Liu, and A. Mushegian, “Evolutionary history of bacteriophages with double-stranded dna genomes,” *Biology Direct*, vol. 2, no. 1, p. 36, 2007.
- [41] S. J. Gould, “The structure of evolutionary theory,” 2002.
- [42] 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](https://doi.org/10.1007/978-1-4613-9586-7_3).
- [43] R. R. Hudson, “Generating samples under a wright–fisher neutral model of genetic variation,” *Bioinformatics*, vol. 18, no. 2, 2002.
- [44] 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.
- [45] D. H. Huson, R. Rupp, and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2010.
- [46] J. Huxley, *Evolution: The Modern Synthesis*. MIT Press, 1942.
- [47] S. O. Jensen and B. R. Lyon, “Genetics of antimicrobial resistance in *Staphylococcus aureus*,” *Future Microbiology*, vol. 4, no. 5, pp. 565–582, 2009.
- [48] 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.
- [49] 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.
- [50] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology*, ser. Applied Mathematical Sciences. Springer, Jan. 2004, vol. 157.
- [51] M. Kreitman, “Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster,” *Nature*, vol. 304, no. 5925, pp. 412–417, Aug. 1983.

- [52] D. M. Kristensen, A. S. Waller, T. Yamada, P. Bork, A. R. Mushegian, and E. V. Koonin, “Orthologous gene clusters and taxon signature genes for viruses of prokaryotes,” *Journal of Bacteriology*, vol. 195, no. 5, pp. 941–950, 2013.
- [53] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. M. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, and Kaul..., “Initial sequencing and analysis of the human genome,” *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

- [54] 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.
- [55] 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.
- [56] M. P. Lesnick, “Multidimensional interleavings and applications to topological inference,” PhD thesis, Stanford University, 2012.
- [57] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, “Comprehensive mapping of long-range interactions reveals folding principles of the human genome,” *Science*, vol. 326, no. 5950, pp. 289–293, 2009.
- [58] 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.
- [59] 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.
- [60] R. MacPherson and B. Schweinhart, “Measuring shape with topology,” *Journal of Mathematical Physics*, vol. 53, no. 7, p. 073516, 2012.
- [61] 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.
- [62] Y. Mileyko, S. Mukherjee, and J. Harer, “Probability measures on the space of persistence diagrams,” *Inverse Problems*, vol. 27, no. 12, p. 124007, 2011.
- [63] D. Morozov, *Dionysus: A C++ library for computing persistent homology*, 2012.
- [64] D. Müllner and A. Babu, *Python mapper: An open-source toolchain for data exploration, analysis and visualization*, 2013.
- [65] M. I. Nelson and E. C. Holmes, “The evolution of epidemic influenza,” *Nature Reviews Genetics*, vol. 8, no. 3, pp. 196–205, Jan. 2007.
- [66] H. C. Neu, “The crisis in antibiotic resistance,” *Science*, vol. 257, no. 5073, pp. 1064–1073, 1992.

- [67] 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 of the United States of America*, vol. 108, no. 17, pp. 7265–7270, 2011.
- [68] 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.
- [69] 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.
- [70] W. H. Organization, “Antimicrobial resistance: Global report on surveillance 2014,” Tech. Rep., 2014.
- [71] 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.
- [72] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, “A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping,” *Cell*, pp. 1–16, 2014.
- [73] H. Rohde, J. Qin, Y. Cui, D. Li, N. J. Loman, M. Hentschke, W. Chen, F. Pu, Y. Peng, J. Li, F. Xi, S. Li, Y. Li, Z. Zhang, X. Yang, M. Zhao, P. Wang, Y. Guan, Z. Cen, X. Zhao, M. Christner, R. Kobbe, S. Loos, J. Oh, L. Yang, A. Danchin, G. F. Gao, Y. Song, Y. Li, H. Yang, J. Wang, J. Xu, M. J. Pallen, J. Wang, M. Aepfelbacher, and R. Yang, “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.
- [74] 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.
- [75] F. Rohwer, M. Youle, and H. Maughan, *Life in Our Phage World*. San Diego, CA: Wholen, Dec. 2014.
- [76] 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.
- [77] B. Schweinhart, “Statistical topology of embedded graphs,” PhD thesis, Princeton University Press, 2015.
- [78] 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.

- [79] 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.
- [80] 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](https://doi.org/10.1089/cmb.2005.12.147).
- [81] 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.
- [82] I. C. 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.
- [83] 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.
- [84] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer, “Fréchet means for distributions of persistence diagrams,” *ArXiv.org*, 2012. arXiv: [1206.2790v2 \[math.ST\]](https://arxiv.org/abs/1206.2790v2).
- [85] J. Wakeley, *Coalescent Theory*. Roberts & Company, 2009.
- [86] 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.
- [87] J. D. Watson and F. H. Crick, “Molecular structure of nucleic acids,” *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.
- [88] A. R. Wattam, D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon, D. Machi, C. Mao, E. K. Nordberg, R. Olson, R. Overbeek, G. D. Pusch, M. Shukla, J. Schulman, R. L. Stevens, D. E. Sullivan, V. Vonstein, A. Warren, R. Will, M. J. C. Wilson, H. S. Yoo, C. Zhang, Y. Zhang, and B. W. Sobral, “Patric, the bacterial bioinformatics database and analysis resource,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D581–D591, 2013.
- [89] C. R. Woese and G. E. Fox, “Phylogenetic structure of the prokaryotic domain: The primary kingdoms,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.
- [90] 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 of the United States of America, vol. 87, no. 12, pp. 4576–4579, 1990.

- [91] 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.
- [92] ———, “Molecules as documents of evolutionary history,” *Journal of Theoretical Biology*, vol. 8, no. 2, pp. 357–366, 1965.