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

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 . . . 2
- 1.2 Evolution as a Topological Space . . . 5
- 1.3 Thesis Organization . . . 9
- 1.4 Original Contributions . . . 10

2 Background 11
- 2.1 Biology . . . 11
  - 2.1.0.1 Genes and Genomes . . . 11
  - 2.1.1 Evolutionary Processes . . . 12
    - 2.1.1.1 Horizontal Gene Transfer . . . 12
  - 2.1.2 Mathematical Models of Evolution . . . 13
    - 2.1.2.1 The Wright-Fisher Model . . . 13
    - 2.1.2.2 The Coalescent Process . . . 14
    - 2.1.2.3 Metrics on Sequences . . . 15
  - 2.1.3 Phylogenetic Reconstruction . . . 15
    - 2.1.3.1 Distance Matrix Methods . . . 16
    - 2.1.3.2 Phylogenetic Networks . . . 16
    - 2.1.3.3 Space of Phylogenetic Trees . . . 16
    - 2.1.3.4 Number of Trees . . . 16
- 2.2 Topological Data Analysis . . . 17
  - 2.2.1 Intuition . . . 18
  - 2.2.2 Mathematical Preliminaries . . . 19
    - 2.2.2.1 Simplicial Complexes . . . 19
    - 2.2.2.2 Homology . . . 20
  - 2.2.3 Constructing Spaces from Real Data . . . 21
    - 2.2.3.1 The Čech and Vietoris-Rips Complexes . . . 21
    - 2.2.3.2 Filtrations . . . 22
  - 2.2.4 Condensed Representations . . . 22

2.2.4.1 The Mapper Algorithm . . . 22
2.2.5 Persistent Homology . . . 23
2.2.5.1 The Persistence Algorithm . . . 25
2.2.5.2 Stability . . . 25
2.2.5.3 Statistical Persistent Homology . . . 27
2.2.5.4 Multidimensional Persistence . . . 28
2.3 Applying TDA to Molecular Sequence Data . . . 28
2.3.1 The Four Gamete Test: The Simplest Example . . . 31
2.3.2 The Space of Trees, Revisited . . . 31

I Theory **33**

3 Quantifying Reticulation Using Topological Complex Constructions **35**
3.1 Introduction . . . 35
3.2 Persistent Homology of Sequence Data . . . 37
3.2.1 Vertical Evolution . . . 37
3.2.2 Reticulate Evolution . . . 38
3.3 Reticulation Quantification Using Homology . . . 39
3.4 Examples . . . 39
3.5 The Median Complex Construction . . . 41
3.5.1 Inclusion . . . 42
3.5.2 Split Decomposition . . . 43
3.6 Interpretation of Higher Dimensional Homology . . . 43
3.7 Čech Complex Construction as an Optimization Problem . . . 44
3.7.1 Molecular Hypothesis . . . 45
3.8 Examples . . . 46
3.8.1 Kreitman Data . . . 46
3.8.2 Buttercup Data . . . 47
3.8.3 Additional Examples . . . 47
3.8.4 Simple Examples . . . 47
3.9 Conclusions . . . 48

4 Parametric Inference using Persistence Diagrams **49**
4.1 Introduction . . . 49
4.2 Warmup: Gaussian Random Fields . . . 51
4.3 The Coalescent Process . . . 51
4.4 Statistical Model . . . 52
4.5 Experiments . . . 55
4.5.1 Coalescent Simulations . . . 55
4.6 Conclusions . . . 56

II Applications: Viruses and Bacteria **57**

5 Bacteriophage Mosaicism **59**
5.1 Introduction . . . 59
5.2 Approach . . . 61
5.2.1 Data . . . 61
5.3 Results . . . 62
5.4 Phage Ecological Properties . . . 63
5.5 Conclusions . . . 63

6 Reassortment in Influenza Evolution **65**
6.1 Introduction . . . 65
6.2 Influenza Virus . . . 66
6.3 Reassortment . . . 66
6.4 Multiscale Flu Reassortment . . . 67
6.5 Prediction of Host Specific Residues . . . 69
6.6 Conclusions . . . 69

7 Reticulate Evolution in Pathogenic Bacteria **71**
7.1 Introduction . . . 71
7.2 Evolutionary Scales of Recombination in the Core Genome . . . 73
7.3 Protein Families as a Proxy for Genome Wide Reticulation . . . 76
7.4 Antibiotic Resistance in *Staphylococcus aureus* . . . 78
7.5 Microbiome as a Reservoir of Antibiotic Resistance Genes . . . 80
7.6 Conclusions . . . 82

8 Prokaryote Reticulate Evolution - Tree of Life **85**
8.1 Introduction . . . 85
8.2 Materials and Methods . . . 85
8.3 Results . . . 86
8.4 Conclusion . . . 86

III Applications: Human Data **87**

9 Human Recombination Rate Mapping **89**

10 Multiscale Topology of Chromatin Folding **93**
10.1 Introduction . . . 93
10.2 Background . . . 95
10.2.1 Long-Range Chromatin Interactions . . . 96
10.2.2 Minimal Cycle Algorithm . . . 97
10.3 Polymer Simulations . . . 98
10.4 Caulobacter Data . . . 98
10.5 Human Data . . . 100
10.6 Conclusions . . . 103

11 Conclusions **105**

Bibliography **107**

List of Figures

1.1 Charles Darwin's Tree . . . 2
1.2 Ford Doolittle's Tree . . . 6
1.3 The coffee mug and the donut . . . 7
1.4 Treelike and reticulate phylogenies . . . 8

2.1 Tree Space . . . 17
2.2 Simplices . . . 19
2.3 Simplicial Complex . . . 19
2.4 . . . 20
2.5 Relationship between chain and cycle and homology. Adapted ("Adapted") from Fasy. . . . 21
2.6 Dimensionality Reduction for EDA . . . 22
2.7 A filtration . . . 24
2.8 The Persistence Pipeline . . . 25
2.9 Applying TDA to Sequence Data . . . 30

3.1 A tree is trivially contractible and has vanishing higher homology. . . . 38

4.1 Two representations of the same topological invariants computed using persistent homology . . . 50
4.2 Distributions of statistics defined on the H_1 persistence diagram for different model parameters . . . 53
4.3 Inference of recombination rate ρ using topological information . . . 55

5.1 Inconsistency of morphological classification in bacteriophage. . . . 61
5.2 Bacteriophage Barcode Diagram . . . 62

6.1 H_1 persistence diagram computed from an avian influenza dataset. . . . 68

7.1 Core genome exchange in *K. pneumoniae* and *S. enterica* . . . 75
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. . . . 76
7.3 Relative recombination rates computed by persistent homology from MLST profile data. . . . 77

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. . 78
7.5 FigFam similarity network of *S. aureus* 80
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. 82

9.1 Calibration 90
9.2 Population Tracks 91

10.1 Three hierarchices of chromatin organization. At the 100 bp scale, DNA chains wrap around protein complexes called nucleosomes. At the megabase scale, these chains are compacted into domains. Lieberman-Aiden *et al.* proposed closed domains form a *fractal globule* structure. At the genome scale, chromosomes fold into the nucleus in separate territories. The fractal globule represents densely packed regions not open for transcription. Fractal globule image from Lieberman-Aiden et al., 2009. Reprinted with permission from AAAS. 96
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. 97
10.3 Polymer Simulation. (A) 50 Mb polymer with 10 fixed loops is allowed to reach an equilbrium conformation. (B) PH identifies 10 H_1 loops. 99
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. 100
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. 101
10.6 Hi-C data chromosome 1 from GM06690 human cell line data, from Lieberman-Aiden et al., 2009. Left: Contact map representation. Right: PH identifies complex multiscale topology. 102
10.7 Distribution of H_1 bar sizes for GM06690 human cell line data shows a bimodal scale of folding patterns. 102

List of Tables

2.1 Dictionary connecting algebraic topology and evolutionary biology 30

3.1 Čech Homology of Hypercube 46

5.1 Phage families defined by the ICTV 60

7.1 Pathogenic bacteria selected for study and sample sizes in each analysis. 74

Chapter 1

Introduction

On the Origin of Species contains a single figure, depicting the ancestry of species as a branching genealogical tree Darwin, 1859 (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 and fill the *universal tree of life*, the set of evolutionary relationships among all extant organisms on Earth Bowler, 2003. Traditionally, this was the realm of phenotype-derived taxonomies [cite]. With the advent of molecular data and computational approaches for tree-inference, evolutionary biology has become a bona fide quantitative discipline. Molecular phylogenetics — tree building — has become the standard tool for inferring evolutionary relationships. Yet a tree is only accurate if the Darwinian model of descent with modification via reproduction is the sole process driving evolution. However, it has long been recognized that there exist alterative evolutionary processes that allow organisms to exchange genetic material through means beyond simple reproduction. Notable examples include species hybridization, horizontal gene transfer in bacteria, and meiotic recombination in eukaryotes. Collectively, these processes are known as *horizontal evolution*, in contrast to descent with modification, an example of *vertical evolution*. Increasing genomic data, powered by new high-throughput sequencing technologies, has shown that these horizontal processes are more prevalent than

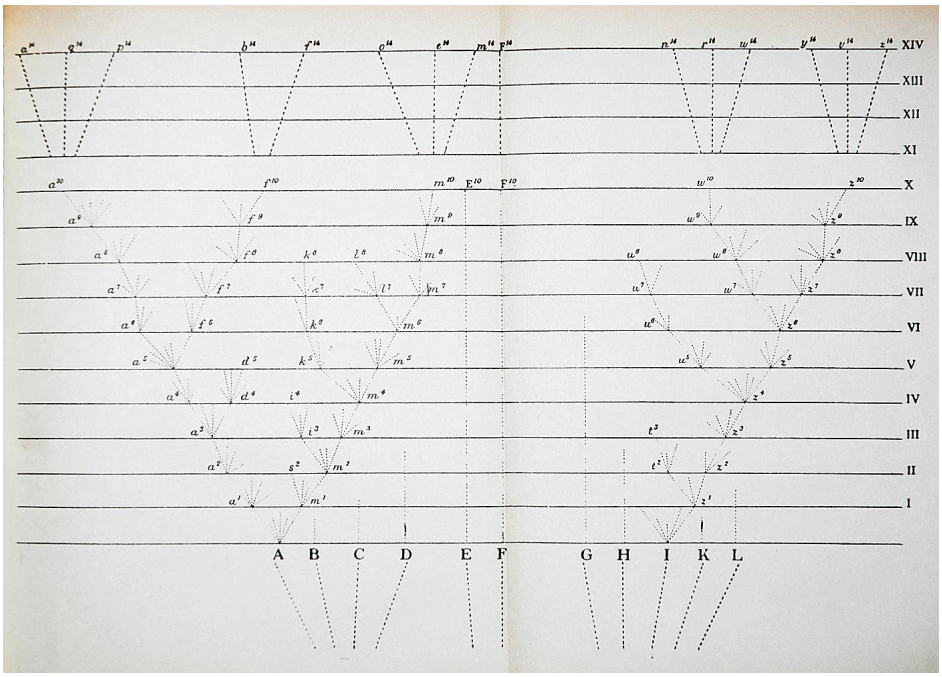


Figure 1.1: The only figure in Darwin's Origin of Species.

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 Doolittle, 1999; O'Malley and Koonin, 2011.

The aim of this thesis is to present a new ways of quantifying and representing nonvertical evolutionary processes using recently developed tools from algebraic and computational topology. These tools fall under the collective heading of *topological data analysis*, a new branch of applied topology concerned with inferring structure in high-dimensional data sets. In the following brief introduction, we survey salient aspects of molecular evolution, the tree paradigm, and the challenges therein. We then introduce the idea of representing evolution as a topological space and give a flavor of our results.

1.1 Molecular Evolution and the Tree Paradigm

The combination of Darwin's ideas of natural selection with Mendelian genetics led to the *modern evolutionary synthesis*, outlined in the first half of the twentieth century in pioneer-

ing works by Fisher, Wright, Haldane, and others (see Huxley, 1942 and Gould, 2002 for historical detail). The modern synthesis was based largely on analysis of distributions of allele frequencies in distinct populations, the purview of classical population genetics. The field was placed on molecular foundations with the Watson and Crick's discovery of the DNA double-helix in 1953 Watson and F. H. Crick, 1953. These developments led to the modern study of *molecular evolution*, the analysis of how processes such as mutation, drift, and recombination act at the cellular level to induce changes in populations and species. While molecular biology has focused on the biochemical and biophysial mechanisms underlying these processes, *molecular phylogenetics* has focused on comparison of macromolecular sequences to infer genealogies and evolutionary relationships. Molecular phylogenetics began with Zuckerkandl and Pauling's recognition that the information encoded in the molecular sequences could be used as a document of evolutionary history in the early 1960's Zuckerkandl and Pauling, 1962; Zuckerkandl and Pauling, 1965. Since that time, the development of numerical approaches for inferring evolutionary relationships has evolved into a mature discipline, with new methods being frequently proposed. The use of molecular sequence data to infer phylogeny is a standard practice across a wide range of biology and ecology.

Two historical results, one from molecular evolution and one from molecular phylogenetics, are worth mentioning. First, Motoo Kimura's neutral theory of evolution, first detailed in 1968 Kimura, 1968 (see Kimura, 1984 for a comprehensive survey). The neutral theory holds that observed genetic diversity is largely a result of genetic drift. At the time the neutral theory was proposed, most biologists assumed that natural selection was the driving force behind genetic diversity. Kimura argued that at the molecular level, the vast majority of mutations are selectively neutral, that is they confer no fitness advantage to the individual organism. With increased sequencing of organisms and populations, tests for selection based on the neutral theory have been developed.

Second, Carl Woese's organization of bacteria, eukarya, and archaea into the three domains of life Woese and Fox, 1977. 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 classes. One of these, which he termed *archaebacteria* was more closely related to eukaryotes than were there the rest of the prokaryotes. This led to the three-domain system of life.

This work had several important consequences. First, it established the use of molecular data to inform about patterns of evolutionary history. Using only morphological data 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 orthologous across all species. Finally, it solidified the tree paradigm as the organizing principle for relating extant species.

However, despite the significant impact this observation had, there remains a subtle difficulty, which Woese himself came 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 Dagan and Martin, 2006). Even more striking, this accounts for less than 0.00005% of the human genome. While recent work has developed approaches for constructing universal trees from larger gene sets Ciccarelli et al., 2006, the fact remains that the vast majority of genomic information is *not* incorporated into the tree.

The reason for this situation is the presence of nonvertical evolutionary forces, as alluded to above. If one were to use a different genomic region to construct atree of species, a different tree topology would be generated. These processes have been well known for some time! 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 Ochman, Lawrence, and Groisman, 2000. 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).

(Expand more of the Doolittle story and the tree paradigm. Species concepts, etc..)

Nonvertical modes of evolution are more than just a theoretical concern: In HIV, frequent recombination confounds our understanding of the early and present epidemics history [cite]. In influenza, gene reassortments lead to antigenic novelty and the emergenence of epidemics [cite]. In several pathogenic bacteria, including E. coli and S. aureus, horizontal gene transfer has been responsbile for the spread of antibiotic resistance [cite].

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. Incompatibilities in the tree model now appear as the rule, not the exception, demonstrating the need for new representations of evolutionary relationships (Doolittle, 1999; Doolittle and Papke, 2006). Many have argued that, in light of genomic evidence of HGT, the very notion of a universal tree of life must be discarded. [cite Doolittle, Koonin]. These and other similar situations, further described below, call for new methods of characterizing evolutionary relationships.

1.2 Evolution as a Topological Space

In this work, we propose the use of new computational techniques, borrowed from the field of algebraic topology, to capture and represent complex patterns of nonvertical evolution. While this may sound obscure, let us unpack the basic idea.

Topology as a field is concerned with properties of spaces that are invariant under continuous deformation. Such properties can include connectedness and the presence of holes, for example. As a paradigmatic example, consider the coffee mug and the donut (Figure 1.3). These two objects are topologically equivalent to $D^2 \times S^1$. That is, there is one connected component and one hole. Algebraic topology quantifies our intuitive notions of shape by associating algebraic structures to different invariants. The most relevant invariants for our

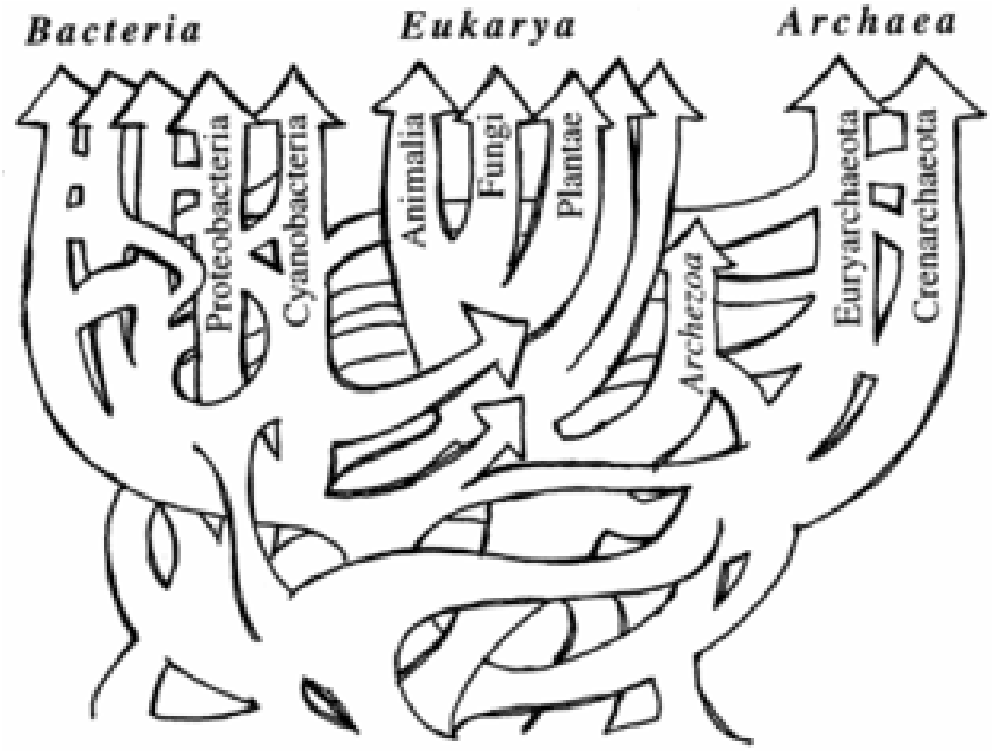


Figure 1.2: W Ford Doolittle's representation of the universal tree of life with nonvertical evolution. (From *Science*, vol. 284, issue 5423, page 2127. Reprinted with permission from AAAS.)

purposes will be the *Betti numbers*, which measure. We give a more complete characterization of Betti numbers in Chapter 2, but a flavor 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, we have $b_0 = 1$, $b_1 = 1$, and $b_n = 0$ for $n > 1$.

Consider again Darwin's branching phylogeny (Figure 1.1) and Doolittle's modified tree accounting for nonvertical processes (Figure 1.2). Imagine these representations as topological spaces. The branching phylogeny has a simple topology, being trivially contractible to a point. In contrast, Doolittle's construction has a much more complex topology, with many holes formed at points where nonvertical processes have occurred. We will adopt a similar

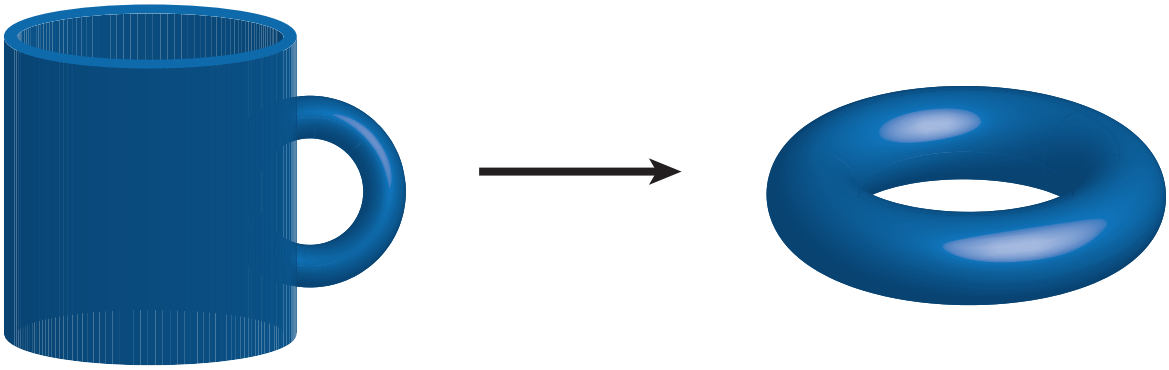


Figure 1.3: The paradigmatic topology example. The coffee mug can be continuously deformed into the donut and are therefore topologically equivalent. Both exhibit $D^2 \times S^1$ topology.

perspective to the coffee mug example and characterize evolutionary spaces as topological spaces using Betti numbers.

To give the very simplest example, consider Figure 1.4. On the left we have a simple rooted three leaf evolutionary topology. We have a single connected component that is trivially contractible, giving $b_0 = 1$ and $b_1 = 0$. 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. Accounting for this generates a single loop, giving $b_0 = 1$ and $b_1 = 1$. The Betti numbers capture the essential difference in the two evolutionary histories.

The remainder of this thesis focuses on expanding upon this idea and applying it to real data sets with the goal of measuring the prevalence and scale of nonvertical evolutionary events. Several immediate questions arise, for instance: 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.

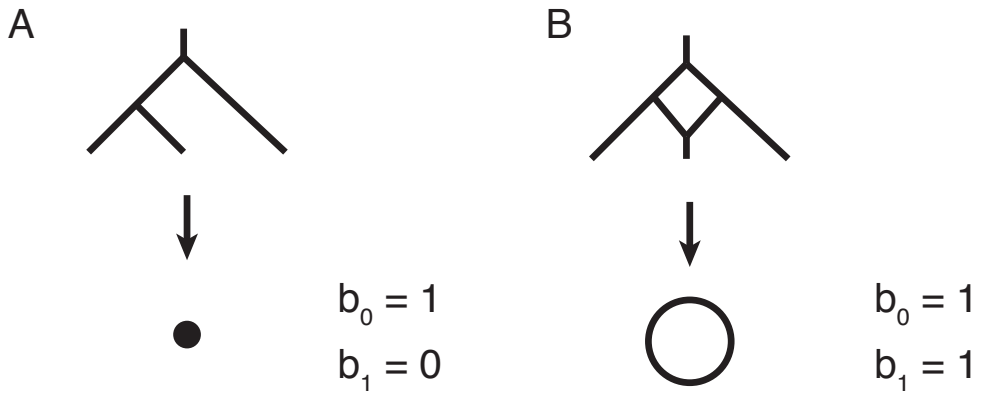


Figure 1.4: (A) A simple treelike phylogeny is contractible to a point. (B) A reticulate phylogeny is equivalent to a circle and not contractible without a cut. The two spaces are no

In this thesis, we use new computational techniques, borrowed from the field of algebraic topology, to capture and represent complex patterns of gene exchange that are obscurbed in current phylogenetic methods. By doing so, we provide a fuller understanding of evolutionary relationships than allowed by current phylogenetic methods. Genomic exchange can be characterized 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. Techniques such as phylogenetic networks and ancestral recombination graphs have been developed to describe reticulate evolution, but they have had only limited success due to difficulties of biological interpretation and computational infeasibility in all but the smallest datasets.

Linkage-based techniques have succeeded in measuring rates of recombination in medium-sized datasets (<200 sequences), but they cannot reveal the scale of these exchanges (i.e., the genetic distance between parental sequences), and they have limited resolution in pinpointing where along a genome such exchanges have occurred. A new mathematical foundation is needed to move beyond these limitations. Genome evolution is an extremely rich subject [cite Genome Architecture book].

1.3 Thesis Organization

The remainder of this thesis is organized as follows.

In Chapter 2 we present background information on the wide range of topics discussed in this thesis. This discussion is chiefly structured into two pieces: (1) background on phylogenetics and population genetics, and (2) background on algebraic topology and the methods of topological data analysis.

In Part I, we develop two complementary approaches for analyzing genomic data using topological data analysis. In Chapter 3, we propose alternative methods of constructing topological complexes that generalize the traditional Vietoris-Rips and Čechcomplexes but are suited to the particular demands of phylogenetic applications. We draw on previous work in phylogenetic networks and use homology to provide a quantitative assessment of reticulation. This work was published in K. Emmett and Rabadan, 2015. In Chapter 4, we develop methods for performing statistical inference using summary statistics contained in the persistence diagram. This is the first such use of persistence diagrams as a tool for performing parametric inference. This work was published in K. Emmett, Rosenbloom, et al., 2014

In Part II we apply our approach to various microorganism datasets. In Chapter 5 we study bacteriophages. In Chapter 6 we study influenza. In Chapter 7 we study pathogenic bacteria and use topological techniques to represent the spread of antibiotic resistance. In Chapter 8 we study prokaryotic evolution and species tree topologies.

In Part III, we apply our approaches to a several problems in human population genetics and biology. 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 toexplore patterns of chromatin folding in the nucleus in both prokaryotic and human datasets. [Hi-C work includes more than human data, and the population structure work will not be included – rename this part?]

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

1.4 Original Contributions

Papers in thesis:

- BIH 2014
- ICML 2014
- Topdrim4bio 1
- Topdrim4bio 2
- Microbes Review
- Cell Systems (submitted)

Additional references:

- Vancouver (PIMS)
- Minnesota (IMA)
- CSHL

Additional papers not in thesis:

- Biofilms / Nature Communications
- Stochseq / BPJ
- Ebola / PLoS Currents
- Mendelian Diseases / Nature Communications

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 intended to motivate the problems discussed in this thesis.

2.1.0.1 Genes and Genomes

The information required to code for biological function is contained in an organisms genome. The genome is a linear string of nucleotides (DNA) that represents the set of genes in an organism. Following the central dogma of biology, DNA is transcribed into RNA, RNA is translated into amino acids, and amino acids are folded into proteins F. Crick, 1970. Proteins comprise the functional unit of biology.

Beyond simply coding for function, the genome includes an imprint of the evolutionary history that gave rise to the organism. By comparing the genomes of multiple organisms,

inferences can be drawn about the evolutionary relationships among extant organisms as well as the processes that generated biological diversity. The field concerned with exploring these relationships is *comparative genomics*.

2.1.1 Evolutionary Processes

There are two main evolutionary processes that have generated the present universe of genomic diversity. These can be conveniently be distinguished as either vertical or horizontal evolution. Vertical, or clonal, evolution is mediated by the accumulation of stochastic mutations over multiple generations. It is vertical evolution that Darwin had in mind when he described the idea of descent with modification, whereby a parent passes genomic information to an offspring subject to random drift. Importantly, vertical evolution will be consistant with a phylogenetic tree model. Horizontal, or reticulate, evolution, refers to a more complex set of processes whereby genomic information can be exchanged between organisms without a parent-offspring relationship. We explain this in more detail below.

Traditionally, evolutionary biology has concerned itself with characterizing relationships in light of vertical evolution alone. However, increasing evidence has pointed to the important role played by horizontal evolution, particularly in prokaryotic evolution.

[Horizontal; reticulate; lateral. Recombination and reassortment. Gene trees vs species trees.]

2.1.1.1 Horizontal Gene Transfer

Horizontal gene transfer refers to a set of processes where genes can be acquired by organisms through means other than reproduction. In prokaryotes, three mechanisms of horizontal transfer are identified, depending on the route by which foreign DNA is acquired Ochman, Lawrence, and Groisman, 2000. 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).

The presence of horizontal gene transfer in a set of organisms can be most clearly identified by comparing the phylogenetic trees built from different genes. If horizontal gene transfer has occurred, the set of *gene trees* will reflect different evolutionary relationships and not be consistent with a single tree topology. A subfield of comparative genomics is concerned with building *species trees* from sets of gene trees, however in the case where there is substantial disagreement among gene trees the very notion of a species tree may be flawed. [Further citations and exposition of Doolittle, Koonin, and Gogarten.]

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. These studies have primarily been replaced by large-scale genomic surveys which have provided unprecedented insight into ancient population structure and historical migrations. [Give an example and cite work of reich / bustamente, etc.]

2.1.2.1 The Wright-Fisher Model

The Wright-Fischer model is a forward time simulation of an evolving population. In the simplest case, the model describes neutral evolution of a constant population size with no structure and constant genome length. The model proceeds in units of generations. At each generation, a member of the population is an offspring of a randomly selected ancestor from the previous generation. This offspring inherits its ancestors genomes, with mutations introduced at some base rate μ. A member of previous generation with no offspring will be considered extinct. [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 Wakeley, 2009. The genealogy is then used to simulate the genetic sequences of the sample. This model is essential to many methods commonly used in population genetics. Starting with a present-day sample of n individuals, each individual's lineage is traced backward in time, towards a mutual common ancestor. Two separate lineages collapse via a coalescence event, representing the sharing of an ancestor by the two lineages. The stochastic process ends when all lineages of all sampled individuals collapse into a single common ancestor. In this process, if the total (diploid) population size N is sufficiently large, then the expected time before a coalescence event, in units of $2N$ generations, is approximately exponentially distributed:

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

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

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

The coalescent with recombination is an extension of this model that allows different genetic loci to have different genealogies. Looking backward in time, recombination is modeled as a splitting event, occurring at a rate determined by population-scaled recombination rate ρ, such that an individual has a different ancestor at different loci. Evolutionary histories are no longer represented by a tree, but rather by an *ancestral recombination graph*. Recombination is the component of the model generating nontrivial topology by introducing deviations from a contractibile tree structure, and is the component which we would like to quantify. Coalescent simulations were performed using `ms` Hudson, 2002.

2.1.2.3 Metrics on Sequences

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 between two aligned sequences.

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. [Worth expanding?]

Jukes-Cantor metric is defined as

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

2.1.3 Phylogenetic Reconstruction

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

Starting with a set of sequences that share some similarity, an alignment is performed. The alignment allows columns of the sequence to be directly compared. 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. (Define additivity – four point condition.) Identifying a pairwise distance matrix with a finite metric space representation is a crucial step that allows most of the machinery described later to be applied.

2.1.3.1 Distance Matrix Methods

Introduced by Cavalli-Sforza and Edwards in 1967 Cavalli-Sforza and A. W. Edwards, 1967 and Fitch and Margoliash in 1967 Fitch and Margoliash, 1967. Compute a matrix of pairwise distances and then find the tree that best approximates those distances. 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 Saitou and Nei, 1987.

2.1.3.2 Phylogenetic Networks

There are several existing methods for representing reticulate evolution. Most of these methods generalize phylogenetic trees into *phylogenetic networks*, which attempt to reconcile the presence of horizontal evolution in sequence data. However, most simply present corrections to phylogenetic trees, which can fail in cases where horizontal evolution is pervasive, as in many prokaryote datasets. Additionally, the resulting neteworks can be complex and difficult to interpret quantitatively. [Expand.]

2.1.3.3 Space of Phylogenetic Trees

Studies of tree space were initiated by Billera, Holmes, and Vogtmann in Billera, Holmes, and Vogtmann, 2001. 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.

2.1.3.4 Number of Trees

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,*

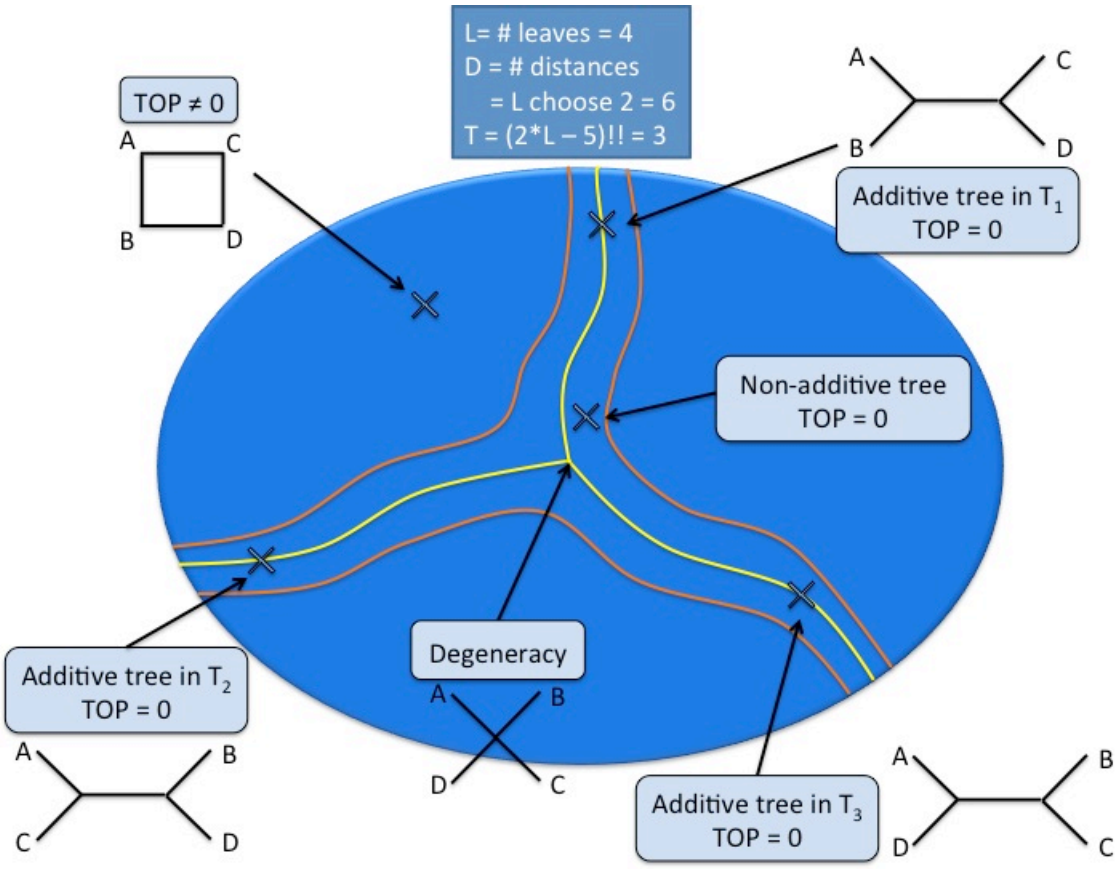


Figure 2.1: Tree Space

more than Avogadro's number of topologies. Phylogeny can be seen as projecting onto the space of trees.

2.2 Topological Data Analysis

Topology is the branch of mathematics that aims to characterize spaces up to deformations. If we take a tree and change the branch lengths, the tree remains a tree. How can we addess if two spaces have the same topology? One trick is to assign to each space some algebraic object (a number, for instance) that does not change under deformation. That object is called a topological invariant. For instance, we can assign to an object the number of loops, which in the case of the mug will be one, and in the case of the tree will be zero. We can only continuously deform spaces with the same invariants. Algebraic topology provides tools

to compute invariants of these spaces.

Real data does not come in the form of perfect continuous spaces, as dealt with by classical topology. Instead, data can be viewed as a high-dimensional point cloud forming a discrete representation of a space. Topological data analysis (TDA) refers to a framework that has been developed in the last 15 years to compute topological properties from finite point clouds, building on developments in computational topology and statistics. Our primary tool is persistent homology, a branch of TDA that computes topological invariants representing multi-scale information about the connectivity and holes in a dataset.

2.2.1 Intuition

Topological Data Analysis (TDA) studies structure in high-dimensional data such as connectedness and the presence of holes. In practice, we observe only a sample of data points, from which we wish to infer an underlying model or generating principle. TDA first builds topological complexes from data, then measures informative properties of these complexes. Persistent homology (PH) is a method from TDA that uses algebraic topology to compute quantitative properties in data, including connectedness and the presence of holes. For excellent review of topological data analysis, see the reviews [Carlsson, Ghrist].

Shape information is indexed by dimension. H_0 information tells us about connected components and is roughly equivalent to a hierarchical clustering. Higher dimensions represent loops (H_1), voids (H_2), and their generalizations in the data, giving a quantitative representation of shape. The topological invariants can be concisely represented ina barcode diagram, which tracks the shape across multiple scale parameters. Each horizontal line in the diagram represents a topological feature.

[Program: encode data as simplicial complex, combinatorial version of a topological space. Properties studied from combinatorial, topological, algebraic perspective.]

Figure 2.2:

Figure 2.3:

2.2.2 Mathematical Preliminaries

Topology: characterize properties of spaces invariant under continuous deformation. Our goal in this section is to get to the point of defining homology.

In this section we give background sufficient to define homology for our purposes, before moving on to applied topology and persistent homology. A more thorough exposition of algebraic topology can be found in Hatcher, 2002.

Associate a collection of algebraic objects with a topological space. Quantify global properties of space. Homotopy and Homology. Homology: properties of chains composed of oriented simplices Elements of homology groups are cycles (chains with vanishing boundary). Two k-cycles are homologous if they differ by the boundary of a (k+1)-chain. Incidence matrix representation...

We provide sufficient background to build a working definition of homology.

2.2.2.1 Simplicial Complexes

Simplex A simplex is something like a point, a line, a triangle, or a higher dimensional generalization. Simplex: generalization of triangle or tetrahedron to arbitraty dimensions. [see Zomorodian]. k-simplex: k-dimensional polytope which is the convex hull of k+1 vertices. Faces. Boundary. See Example.

Simplicial Complex Glue together simplices such that the following holds:

1. Any face of a simplex in K is also in K

2. The intersection of any two simplices in K is a face of both

Figure 2.4:

Chains, cycles and boundaries Boundary operator $\partial_k : C_k \to C_{k-1}$. Action of boundary operator on a simplex σ is defined as:

$$\partial_k \sigma = \sum_i (-1)^i [v_0, v_1, ..., \hat{v}_i, ..., v_n]. \tag{2.3}$$

A chain $C \in C_k$ is called a cycle if $\partial_k C = 0$. Chain with empty boundary. Set of cycles forms a group.

Boundary operator defines a chain complex C_*:

$$\ldots \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} ... \xrightarrow{\partial_2} C_n \xrightarrow{\partial_1} C_{n-1} \xrightarrow{\partial_0} 0 \tag{2.4}$$

Important property:

$$\partial_{k-1} \partial_k = 0 \forall k \tag{2.5}$$

Intuitively, a boundary has no boundary.

C is the set of The k-th cycle group is $Z_k = \ker \partial_k$. Z_k defines the set of all cycles of dimension k. B_k defines the set of all boundaries of dimension k. That is, elements of B_k serve as boundaries of $(k+1)$-chains.

TDA+PH: number and type of holes. which holes are essential and which are unimportant.

Get to the point where can define homology. Chain complex. Boundary operators Work only over mod 2 Homology (0,1) coefficients. Torsion observed in the image patch data set, but no reason to think it is present biologial data sets we examine.

2.2.2.2 Homology

Abelian groups generated by holes in the space. Betti number is the rank of the homology group.

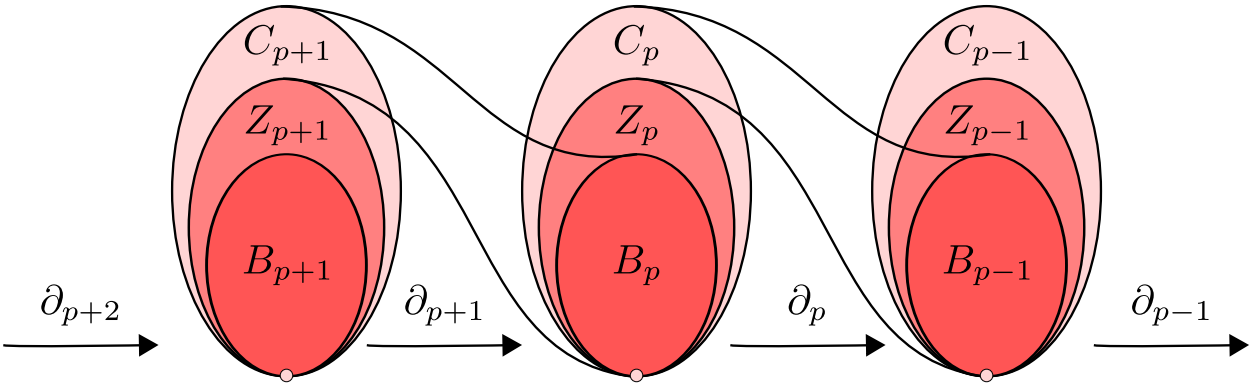


Figure 2.5: Relationship between chain and cycle and homology. Adapted ("Adapted") from Fasy.

Simplical homology. Abelian groups. First homology group: abelianization of the fundamental gorup. Quotient group

Recalling out definition of the boundary and cycle groups, define a quotient group

$$H_k = Z_k / B_k = \ker \partial_k / \operatorname{im} \partial_{k+1} \tag{2.6}$$

Closed chains (cycles) up to boundary of higher dimensional cycles. Elements are classes of homologous cycles.

2.2.3 Constructing Spaces from Real Data

In practice we have real data which comes to us as points in a high-dimensional space. 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 Metric space with a finite number of points. In topological data analysis, our spaces of interest are finite metric spaces.

2.2.3.1 The Čech and Vietoris-Rips Complexes

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

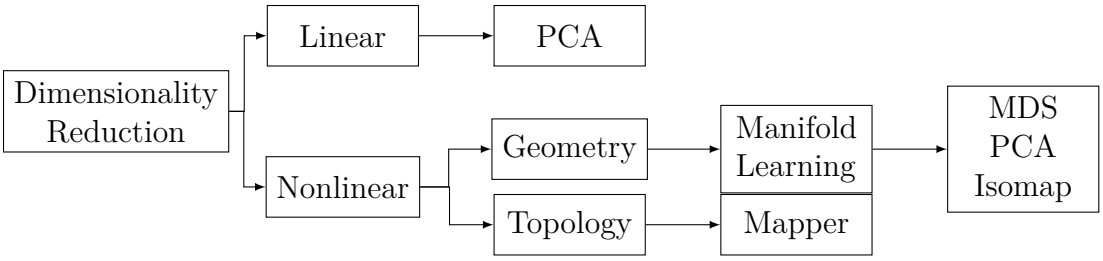


Figure 2.6: Dimensionality Reduction for EDA

$$\cap_i B(v_i, \epsilon) \neq 0 \tag{2.7}$$

The Vietoris-Rips complex is defined as

$$\text{VietorisRips}(r) = \{\sigma \in S \mid \text{diam}(\sigma) \leq 2r\} \tag{2.8}$$

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

2.2.3.2 Filtrations

A series of inclusions.

2.2.4 Condensed Representations

2.2.4.1 The Mapper Algorithm

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

Mapper algorithm was first developed in Singh, Mémoli, and Carlsson, 2007. Mapper is coordinate free and depends onnly on the similarity of points as measured by the distance function. Further exposition can be found in Lum et al., 2013. Used in biology example in Nicolau, Levine, and Carlsson, 2011 top classify breast cancer subtypes. In our work we use the commercial Mapper implementation Ayasdi Inc., 2015. An open-source implementation

of the Mapper algorithm is available in the Python Mapper package Müllner and Babu, 2013.

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

2.2.5 Persistent Homology

Persistent refers to capturing structure that extends over multiple spatial resolutions and is in some sense robust, while homology refers to the type of topological structure that is being computed.

How to extend homology to finite metric spaces? Data -> Sets of complexes -> Vector spaces

We summarize persistent homology from the perspective of an end-user. For detailed background, see the reviews Carlsson, 2009; Ghrist, 2008 and the books Edelsbrunner and Harer, 2010; Zomorodian, 2005. In brief, persistent homology computes topological invariants representing information about the connectivity and holes in a dataset. A dataset, $S = (s_1, \ldots, s_N)$, is represented as a point cloud in a high-dimensional space (not necessarily Euclidean). From the point cloud, a nested 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.

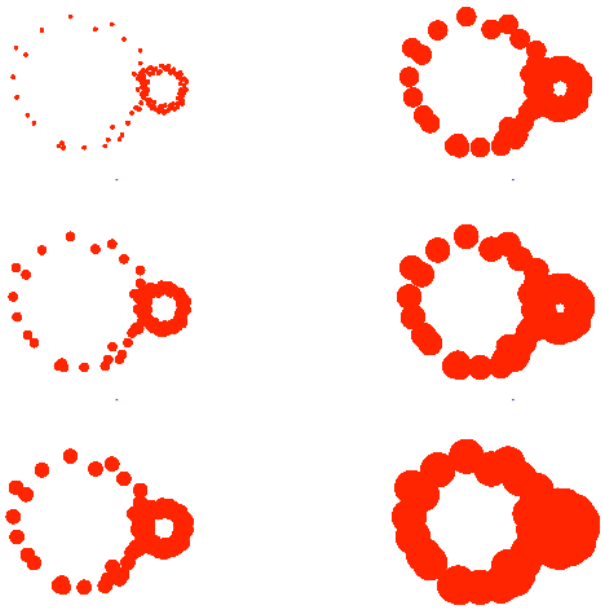


Figure 2.7: Example of constructing a filtration.

The topological invariants in the filtration can be concisely represented in a barcode diagram, a set of line segments ordered by filtration value on the horizontal axis. In the barcode diagram,

The topological invariants in the filtration can be concisely represented in a barcode diagram, a set of line segments ordered by filtration value on the horizontal axis (Figure XXX).

Invariants can be equivalently represented by a persistence diagram, a scatter plot with the birth time on the horizontal axis and the death time on the vertical axis.

The intuition behind persistent homology is that is that good or interesting features will persist over longer scales That is, they will be more robust. In the barcode diagram this corresonds to long bars, and in the persistence diagram, this corresponds to points far from the diagonal. Invariants that persist for only short scales are likely to be noise or artifacts of incomplete sampling. The questions of how to rigorously determine what makes a good interval is an open question that is currently being addressed by a number of different groups. We discuss this further in Section 2.2.5.3.

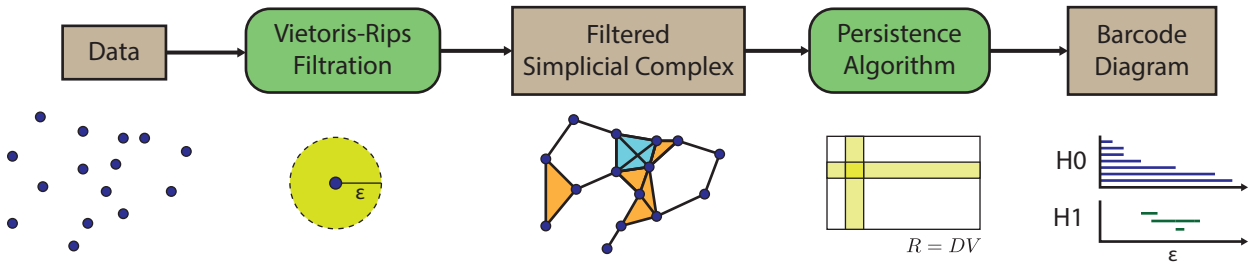


Figure 2.8: The Persistence Pipeline.

2.2.5.1 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 represetnative 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 simplciial set is often the limiting factor here, they collapse simplicies 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 Z_2.

Several packages for computing persistent homology have been developed [Dionysus, Javaplex, Guidi, phom] and TDA frontend for R. Persistent homology is computed using Dionysus Morozov, 2012.

2.2.5.2 Stability

An important aspect of peristent 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 $D \to D'$, what can I say about the subsequent change in barcodes $B \to B'$? 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 (Chazal, Cohen Steiner, et al., 2009). 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.

$$d_H(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(X,Y)\right\} \tag{2.9}$$

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.

$$d_{GH}(X,Y) = \inf_{f,s} d_H(X,Y) \tag{2.10}$$

Next, we consider how to define the distance between two persistance 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

$$d_\infty(a,b) = \max\{|a_x - b_x|, |a_y - b_y|\}. \tag{2.11}$$

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*.

$$d_B(A,B) = \inf_{n:A\to B}\sup_{x\in X} ||x - \eta(x)||_\infty \tag{2.12}$$

The result of Chazal, Cohen Steiner, et al. (2009) 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.*

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

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

2.2.5.3 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 (Fasy et al., n.d.). 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 Bubenik and Kim, 2007; Bubenik, 2015. 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 Turner et al., 2012 and Mileyko Mileyko, Mukherjee, and Harer, 2011.

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

Wasserstein distance between diagrams.

$$W_p(D_1, D_2) = \left(\inf_{\gamma} \sum_{x \in D_1} \|x - \gamma(x)\|_\infty^p \right)^{1/p} \tag{2.14}$$

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

2.2.5.4 Multidimensional Persistence

First laid out in Carlsson and Zomorodian, 2009. More work in Lesnick, 2012. 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 betwee genomes. This defines our finite metric space. From there, methods

from TDA such as persistent homology and mapper can be applied. Phylogenetic information can be read off the resulting topological structures.

An important foundational point was described in Chan, Carlsson, and Rabadan, 2013. 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 2.9. Consider the reticulate phylogeny (Figure 2.9A): 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 2.9B, 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 2.9C we demonstrate the construction of a filtered simplicial complex, showing how the connectivity changes as the scale parameter ϵ is increased. Finally, in Figure 2.9D 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 higlighted sequences. The scale over which this bar persists represents the amount of evolutionary time separating the parents and the reticulate offspring. Additionally, the persistence algorithm will return a generating basis for a particular homology group, which we can use to identify the particular mixtures of sequences involved a reticulation. In this way, we can analyze both the scale and frequency

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

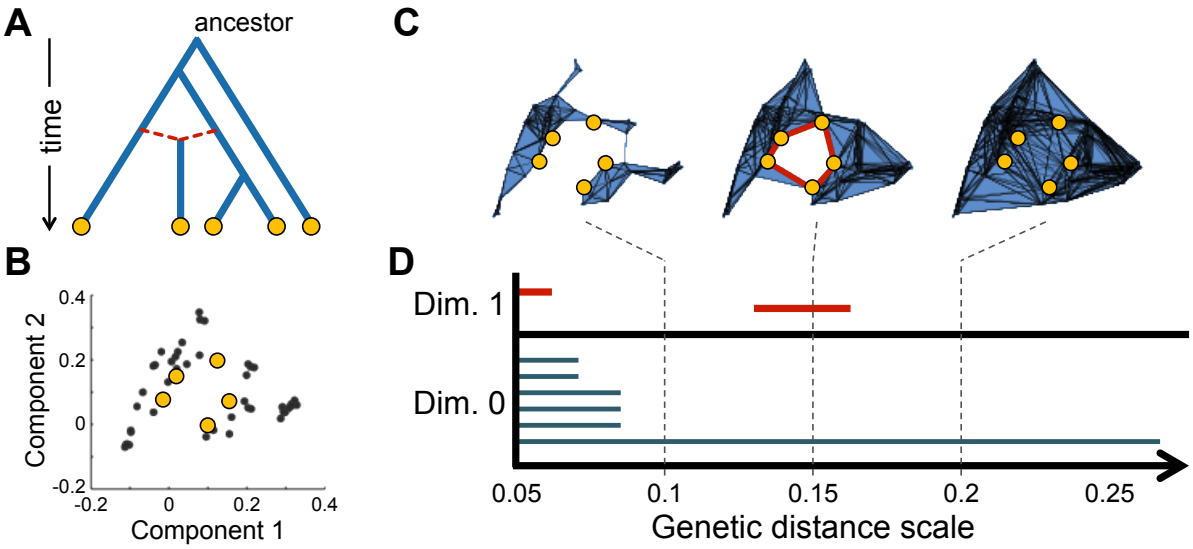


Figure 2.9: Applying persistent homology to genomic data. (A) An evolutionary genealogy including reticulation. (B) Data projected into 2-dimensions. (C) Construction of a filtered simplicial complex. (D) The resulting multiscale barcode diagram.

of reticulation in genomic data sets.

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

2.3.1 The Four Gamete Test: The Simplest Example

In population genetics and phylogenetics, the four gamete test is a simple method for testing for the presence of reticulate evolution. In any given dataset, the simultaneous presence of haplotype patterns 00, 10, 01, and 11 is incompatible with strictly clonal evolution. [Needs more explanation why?]

Using a Hamming metric, we have a finite metric space which describes a simple diamond. The barcode diagram will reflect nontrivial topology in the interval $[1, 2)$. This simple example serves as motivation for much of the work that follows.

2.3.2 The Space of Trees, Revisited

In Section XX we introduced the tree space as a subspace of the space of finite metric spaces. Real data will not often sit in tree space, and the goal of phylogenetics can be seen as finding the best tree representation of data via some projection onto tree space. Our program can be interpreted in a similar fashion: data comes to us as a finite metric space. Rather than projecting onto tree space, we want to characterize it as it is, by defining topological invariants which characterize the deviation from tree-like additivity. The further the data sits from tree space, the more reticulation we expect to find.

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 (Chan, Carlsson, and Rabadan, 2013), where recombination rates in viral populations were estimated by computing L_p norms on barcode diagrams. In that paper, it was shown that persistent homology provides an intuitive quantification of reticulate evolution in sequence data by measuring deviations from tree-like additivity. Reticulation is manifest as nonvanishing higher homology ($H_n > 0$ for $n > 0$) in the filtration. Using persistent homology asa tool to measure reticulate evolution is useful because it

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

of tracking the scale of reticulate events.

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

Due to the coarseness of the distance filtration, only those reticulations which have sufficiently strong support in the sequence data will be detected. By coarseness, we mean that thae distance filtration... Small distortions in the metric space, due possibly to incomplete population sampling or weakly supported reticulations, will reduce sensitivity. Looking to increase the resolution of our approach led us to consider a class models which construct a *median graph* from a set of sequences. Median graphs form the basis for a large number of phylogenetic network algorithms and have been extensively studied over the past several decades (Bandelt, Forster, and Röhl, 1999). The approach is closely related to split decomposition, and it can be shown that the objects resulting from the two methods are identical (Bandelt and A. W. Dress, 1992). The median graph approach imputes putative evolutionry ancestors into the set of vertices, and forms a network representing the incompatibile splits present in the sequence data. A common task has been to quantify the complexity of the resulting network. We show that a filtration of complexes built from the median graph vertex set is a fast an efficient way to characterize the complexity of a phylogenetic network. Due to a result of Gromov, we know that the complexes built on this vertex set will be cubical, making the barcode diagram simple to interpret Gromov (1987).

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

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

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

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

3.2 Persistent Homology of Sequence Data

In this section we briefly review the ideas in Chan, Carlsson, and Rabadan (2013) as they relate to the application of persistent homology to sequence data.

3.2.1 Vertical Evolution

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

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

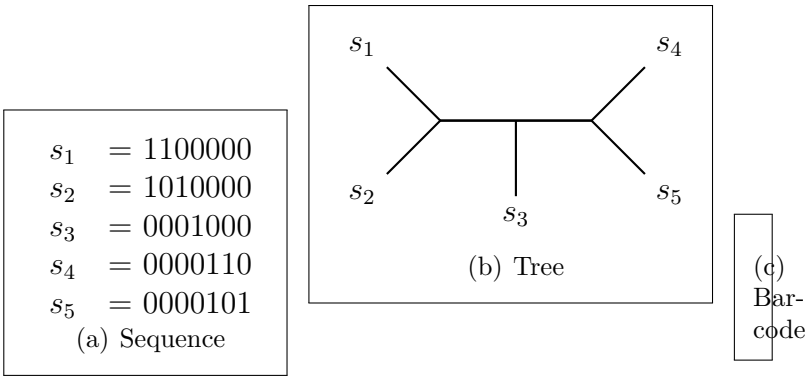


Figure 3.1: A tree is trivially contractible and has vanishing higher homology.

on the points such that

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

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

A tree is trivially contractible, and hence has vanishing higher homology. This result was proven for sequence data in (Chan, Carlsson, and Rabadan, 2013). In practice, most data is not additive. The field of phylogenetics is essentially tasked with finding the *best* tree given sequence data, for some notion of best.

3.2.2 Reticulate Evolution

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

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

is incompatible with strictly vertical evolution in an infinite sites model. It provides direct evidence for reticulate evolution. One way to quantify recombination in a set of sequences is the Hudson-Kaplan test, which counts the minimum number partitions required in the data such that within each partition the all sites are compatible (Hudson and Kaplan, 1985).

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 (2005). 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 (2005) 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,

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

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

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

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

Filtrations on Buneman graphs have been defined previously (A. Dress, Huber, and Moulton, 1997), but not using an explicit sequence representation. The filtration defined in A. Dress, Huber, and Moulton (1997) is based on a complicated polytope construction scheme defined directly from the split decomposition. Given that all median graphs are split networks (Huson, Rupp, and Scornavacca, 2010), 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 in-

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

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

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

3.5.2 Split Decomposition

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

3.6 Interpretation of Higher Dimensional Homology

In Chan, Carlsson, and Rabadan (2013) it was argued that higher dimensional homology (H_d for $d > 1$) is evidence for 'more complex' reticulate events. Here we try to make this notion more precise, showing by way of examples that higher dimensional homology can be interpreted as evidence of multiple interacting reticulate events. First, we detour slightly and introduce a Čech complex construction that will increase our sensitivity to these events.

3.7 Čech Complex Construction as an Optimization Problem

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

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

where $B_x(r)$ is the ball of radius r centered at vertex x. By the nerve lemma, the homotopy type of the Čech covering is guaranteed to be identical to that of the original topological space (Borsuk, 1948).

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.*(Gärtner, 1999) 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

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

problem in L_1 as a generic convex optimization problem, and use standard library solver. That is, we define a $d+1$ dimensional optimization problem where x is the miniball center and R is the miniball radius.

The problem is stated as

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

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

3.7.1 Molecular Hypothesis

Gromov proved that a median graph is the 1-skeleton of a CAT(0) cubical complex (Gromov, 1987). The homology of a cubical complex can be efficiently computed using the methods of Kaczynski, Mischaikow, and Mrozek (2004) through a slightly different construction. We define a cubical flag complex and build a filtration dimension by dimension (to expand on this point...) The barcode diagram will then have the natural interpretation of being composed of sets of hypercubes of varying dimension. If we consider each bar of dimension n in the barcode diagram in turn, we can determine the incompatible sites that it represents. Dimension 1 bars (2-cubes) will have one pair of incompatible sites with four haplotypes. Dimension 2 bars (3-cubes) will have three pairs of incompatible sites with eight haplotypes. In general, n bars will represent $n+1$-cubes in which all $2^{(n+1)}$ haplotypes are present in the vertices of the generating cycle.

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

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

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

Table 3.1: Čech Homology of Hypercube

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

3.8 Examples

3.8.1 Kreitman Data

A benchmark dataset in recombination studies is the Kretiman data (Kreitman, 1983). The dataset consists of eleven sequences (nine unique) of the Adh locus from *Drosophilia melanogaster* collected from various locations, with 43 segregating sites. Several methods have been applied to this data to estimate the minimum number of recombinations present in this data. The Hudson-Kreitman test yields 6, while Song-Hein computed 7. The persistent

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

3.8.2 Buttercup Data

3.8.3 Additional Examples

See Huson, Rupp, and Scornavacca (2010).

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 Fasy et al., n.d.; Blumberg et al., 2014; Chazal, Glisse, et al., 2014. 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

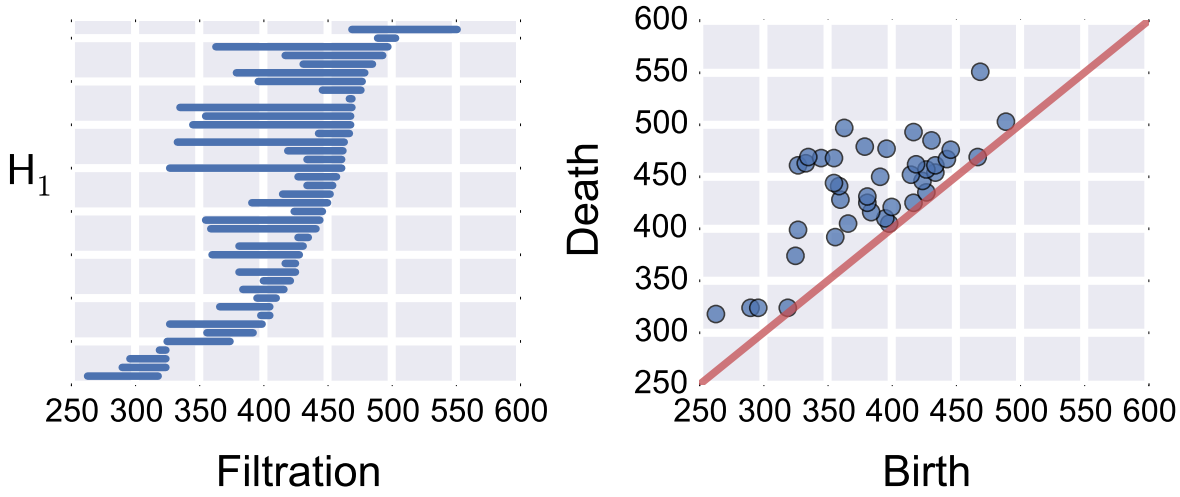


Figure 4.1: Two representations of the same topological invariants, computed using persistent homology. Left: Barcode diagram. Right: Persistence diagram. Data was generated from a coalescent simulation with $n = 100$, $\rho = 72$, and $\theta = 500$.

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 Adler et al., 2010. 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 Wakeley, 2009. The genealogy is then used to simulate the genetic sequences of the sample. This model is essential to many methods commonly used in population genetics. Starting with a present-day sample of n individuals, each individual's lineage is traced backward in time, towards a mutual common ancestor. Two separate lineages collapse via a coalescence event, representing the sharing of an ancestor by the two lineages. The stochastic process ends when all lineages of all sampled individuals collapse into a single common ancestor. In this process, if the total (diploid) population size N is sufficiently large, then the expected time before a coalescence event, in units of $2N$ generations, is approximately exponentially distributed:

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

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

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

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

The coalescent with recombination is an extension of this model that allows different genetic loci to have different genealogies. Looking backward in time, recombination is modeled as a splitting event, occurring at a rate determined by population-scaled recombination rate ρ, such that an individual has a different ancestor at different loci. Evolutionary histories are no longer represented by a tree, but rather by an *ancestral recombination graph*. Recombination is the component of the model generating nontrivial topology by introducing deviations from a contractibile tree structure, and is the component which we would like to quantify. Coalescent simulations were performed using `ms` Hudson, 2002.

4.4 Statistical Model

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

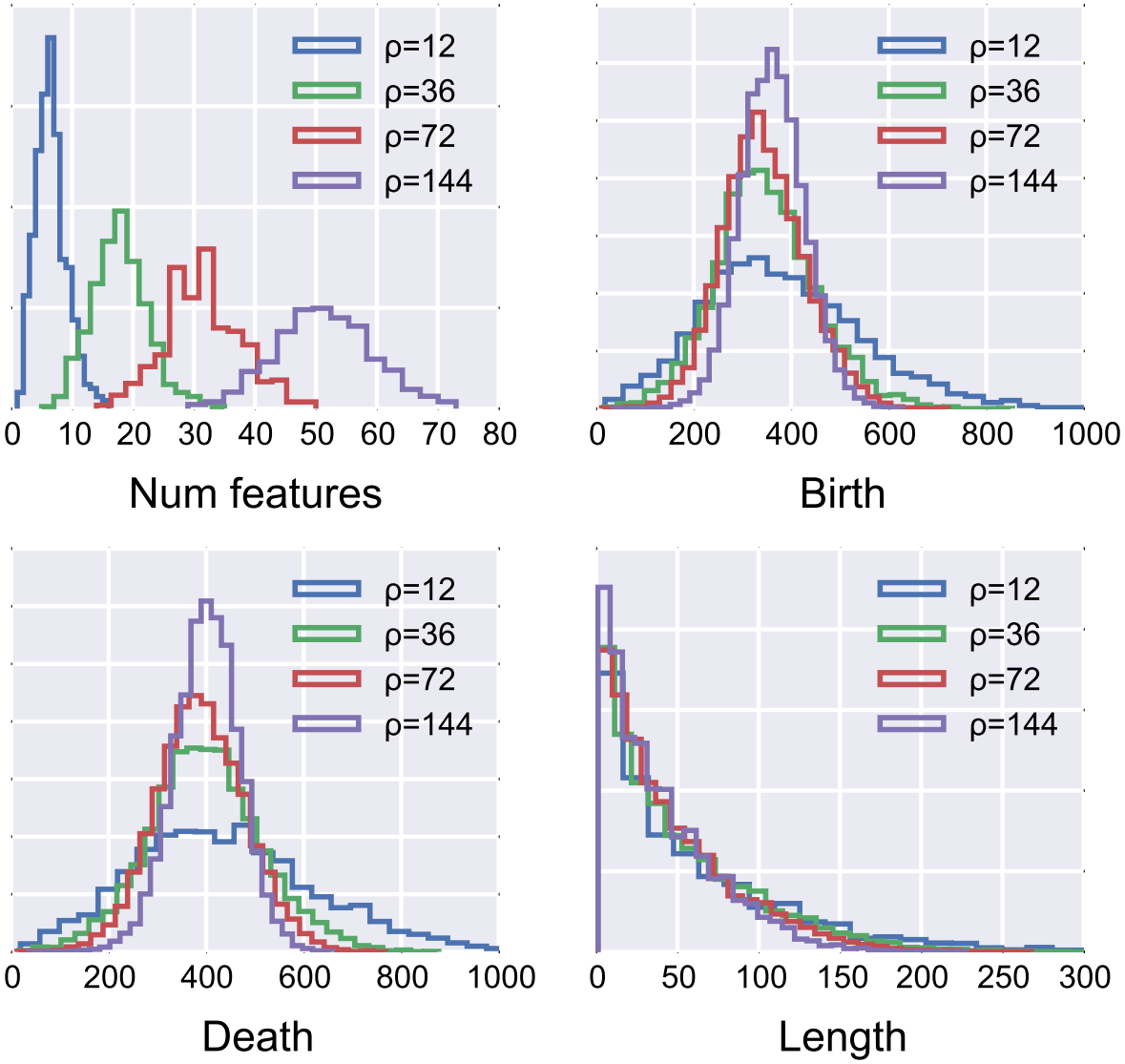


Figure 4.2: Distributions of statistics defined on the H_1 persistence diagram for different model parameters. Top left: Number of features. Top right: Birth time distribution. Bottom left: Death time distribution. Bottom right: Feature length distribution. Data generated from 1000 coalescent simulations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

persistence lengths are independent of ρ.

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

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

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

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

Birth times are Gamma distributed with shape parameter

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

and scale parameter

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

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

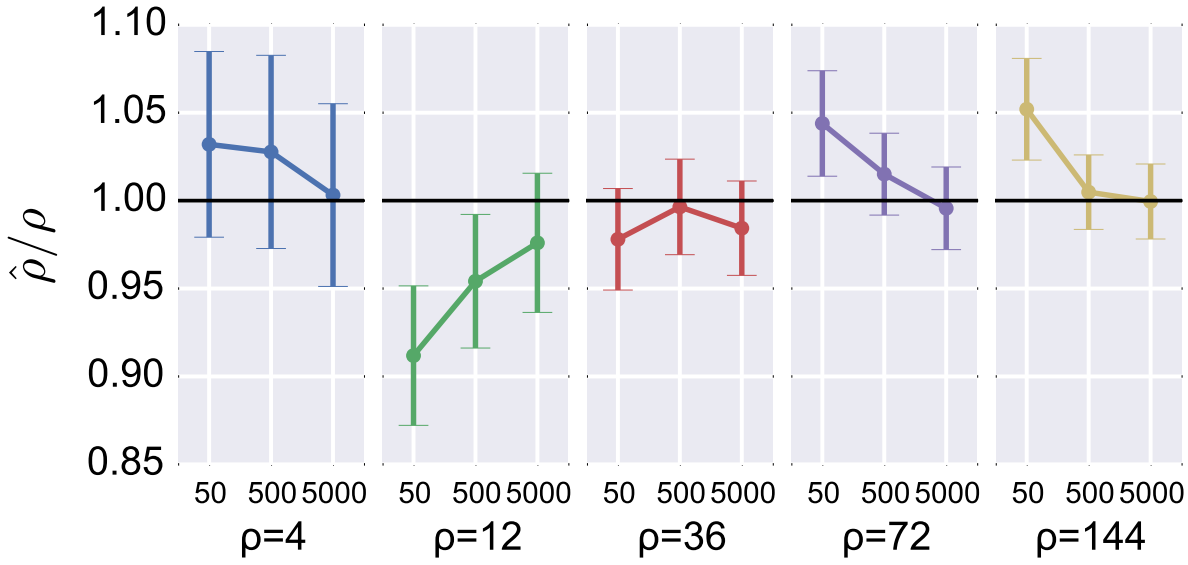


Figure 4.3: Inference of recombination rate ρ using topological information. The recombination rate ρ is estimated for five values $\{4, 12, 36, 72, 144\}$ at three different mutation rates $\{50, 500, 5000\}$. Mean estimate over 500 simulations and 95% confidence interval is shown.

4.5 Experiments

4.5.1 Coalescent Simulations

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

4.6 Conclusions

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

Part II

Applications: Viruses and Bacteria

Chapter 5

Bacteriophage Mosaicism

5.1 Introduction

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 Taxonomy of Viruses, 2012. 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. Nucleic acid composition is either double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), or single-stranded RNA (ssRNA). Of these dsDNA is by far the most common. 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 extremeley 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

Table 5.1: Phage families defined by the ICTV

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

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 noncontracile phages. P22 belongs to the Podoviridae family comprised of short tail phages. Visually, it appears that HK97 and L5 should indeed be classified as distinct from P22. However, genomic analysis 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 Rohwer and R. Edwards, 2002. 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-

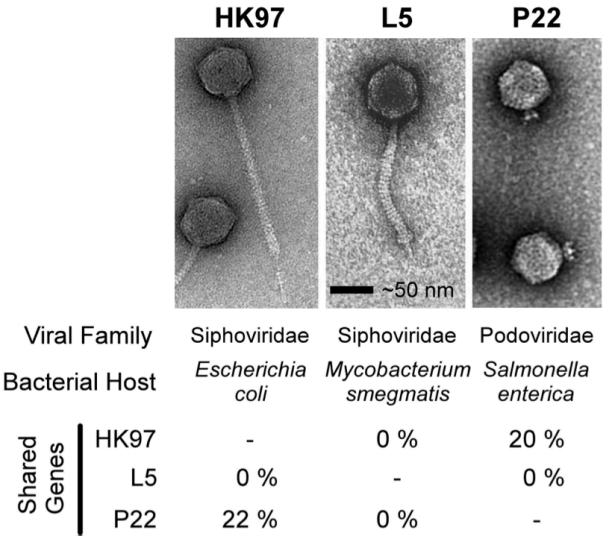


Figure 5.1: Inconsistency of genomic and morphological approaches. HK97 and L5 are classified under same viral family, despite sharing no homologous genetic content. P22 on the other hand shares many genes. Figure adapted from (Lawrence, Hatfull, and Hendrix, 2002)

Mendea *et al.* Lima-Mendez et al., 2008 and Kristensen *et al.* Kristensen et al., 2013. 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.

5.2 Approach

5.2.1 Data

First data set follows that from Lima-Mendez et al., 2008 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: nxp binary gene presence/absence matrix. 29 outlier phages, discard and keep only subset S277.

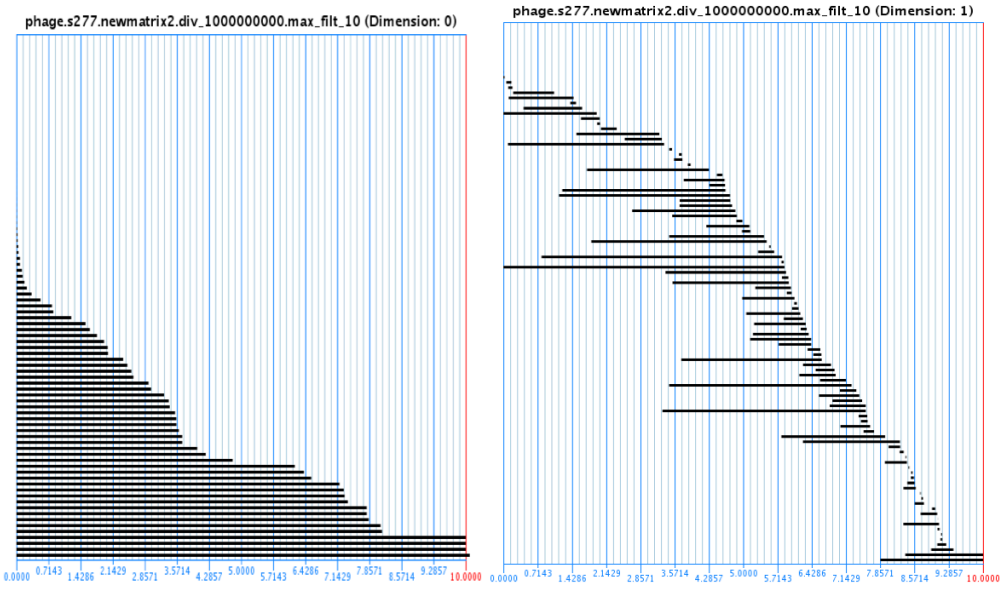


Figure 5.2: Bacteriophage Barcode Diagram using the Lima-Mendez dataset.

Second data set follows that from Kristensen et al., 2013 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 Glazko et al., 2007.

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 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.

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 recombination. 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.

6.2 Influenza Virus

Influenza is a single-stranded RNA virus that is naturally found in avian populations. Each viral genome has eight genetic segments. [More background info.]

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.

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 coinfected with two different viral strains, reassortment of these segments can occur, such that viral offspring is a genetic mixture of the two parental strains. Reassortment is of substantial medical interest, and has been connected with the outbreak of influenza epidemics.

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 6.1, along with density estimates for the birth and death distributions. Both birth and death times appear strongly bimodal, unlike in the coalescent simulations, which were strictly unimodal. This suggests two distinct scales of topological structure. Using the representative cycles output by Dionysus on a subset of this data, we classified features as intrasubtype (involving one HA subtype) and intersubtype (involving multiple HA subtypes). The H_1 barcode diagram for this data is shown in the Figure 6.1 inset. Intrasubtype features, in blue, occur at an earlier filtration scale than intersubtype features, in green. The multiscale topological approach of persistent homology can distinguish biological events occuring at different genetic scales.

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

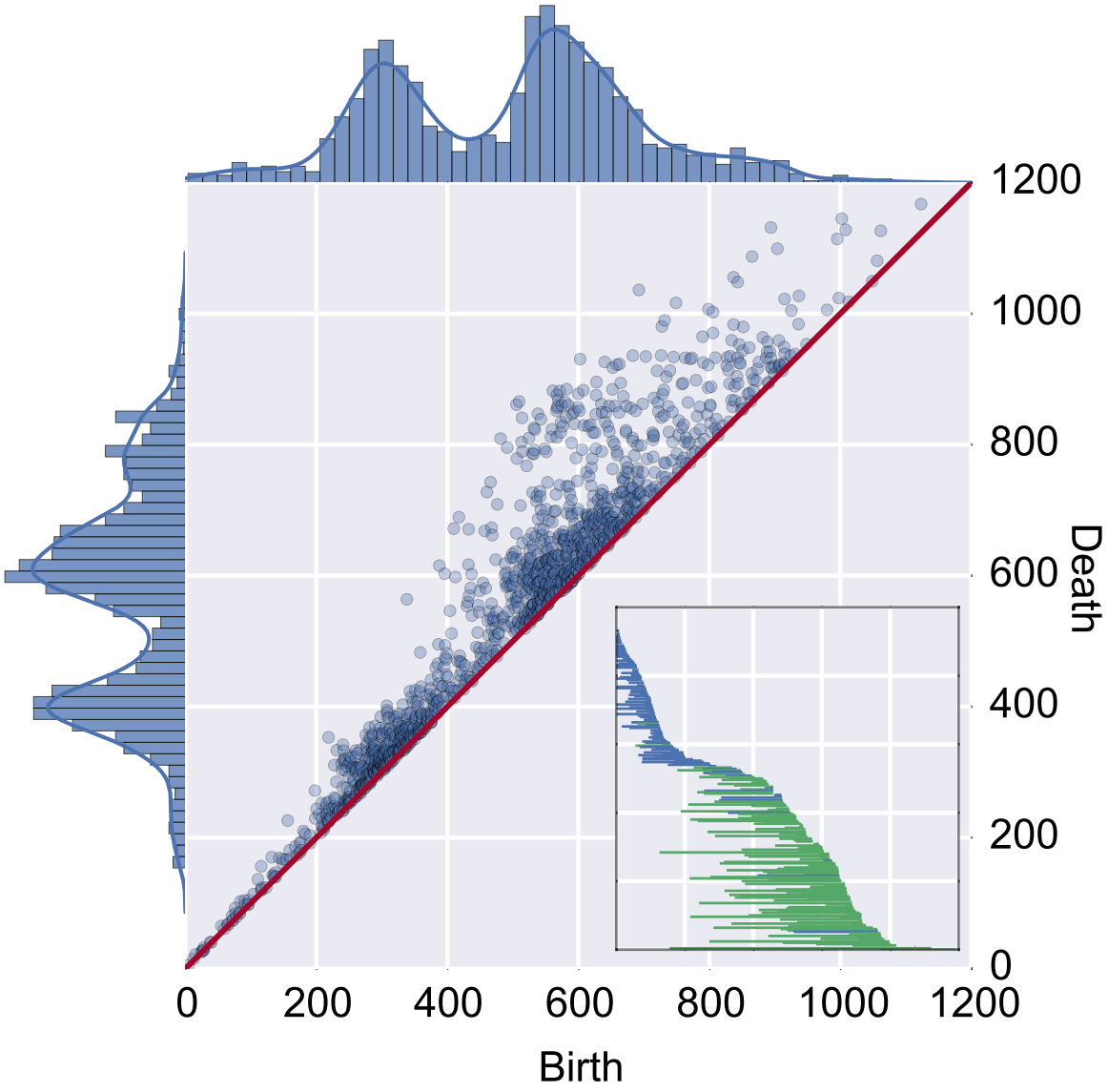


Figure 6.1: The H_1 persistence diagram computed from an avian influenza dataset. On the top and left are plotted the marginal distributions of birth and death times, along with a density estimate for each distribution. The bimodality indicates two scales of topological structure. Inset: The barcode diagram for a subset of this data. Blue bars have representative cycles involving only one subtype, green bars have cycles involving multiple subtypes.

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

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 Neu, 1992; Thomas and Nielsen, 2005; WHO, 2014. 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 Ochman, Lawrence, and Groisman, 2000. 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 Rohde et al., 2011. 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 Woese and Fox, 1977. 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 Dagan and Martin, 2006. 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 Fraser, Hanage, and Spratt, 2007. 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 WHO, 2014. Using persistent homology, we characterize the rate and scale of recombination in the core genome using multilocus sequence data. To extend our characterization to the whole genome, we use protein family annotations as a proxy for sequence composition. This allows us to compute a similarity matrix between strains. Comparing persistence diagrams gives us information about the relative scales of gene transfer at arbitrary loci. The species selected for study and the sample sizes in each analysis are specified in Table 7.1. Next, we explore the spread of antibiotic resistance genes in *S. aureus* using Mapper, an algorithm for partial clustering and visualization of high dimensional data Singh, Mémoli, and Carlsson, 2007. 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 resistence as an example. In this environment, benign bacterial strains can harbor known resistance genes. We use a network analysis to visualize the spread of antibiotic resistance gene *mecA* into nonnative phyla. Each indidivual has a unique microbiome, and we speculate that microbiome typing of this sort may useful in developing personalized antibiotic therapies. These results suggest an important role for topological data mining of -omics scale data in clinical applications and personalized medicine.

7.2 Evolutionary Scales of Recombination in the Core Genome

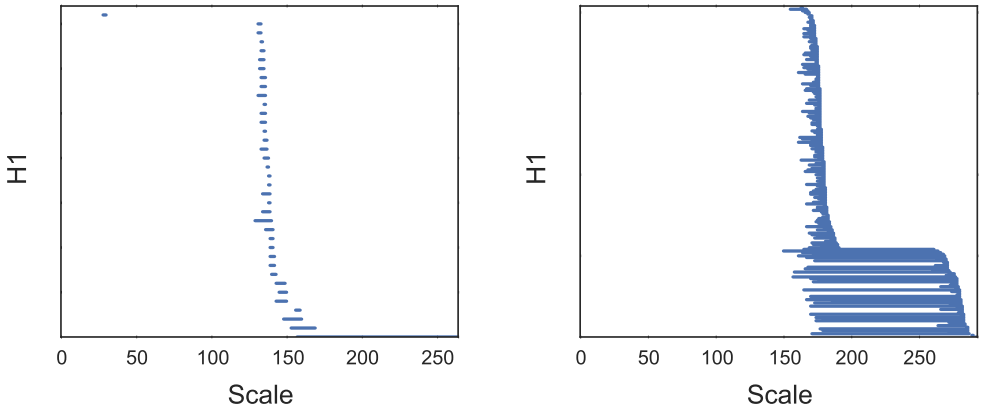
Multilocus sequence typing (MLST) data was used to examine scales of recombination in the core bacterial genome. MLST is a method of rapidly assigning a sequence profile to a sample bacterial strain. For each species, a predetermined set of loci on a small number of housekeeping genes are selected as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

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

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

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

We investigate genetic exchange in the twelve pathogens using MLST data from PubMLST Jolley and Maiden, 2010. 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 $\nu = 0$ de Silva and Carlsson, 2004. The computation is performed using javaplex Tausz, Vejdemo-Johansson, and Adams, 2014. An example of our



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

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

output is shown in Figure 7.1, where we plot the H_1 barcode diagrams for *K. pneumoniae* and *S. enterica*. The two species have distinct recombination profiles, characterized by the range of recombinations: *K. pneumoniae* recombines at only one short-lived scale, while *S. enterica* recombines both at the short-lived scale and a longer-lived scale. We repeat this analysis for each species, and plot the results as a persistence diagram in Figure 7.2. Among the bulk of pathogens there appears to be three major scales of recombination, a short-lived scale at intermediate distances, a longer-lived scale at intermediate distances, and a short-lived scale at longer distances. *H. polyori* is a clear outlier, tending to recombine at scales significantly lower than the other pathogens.

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

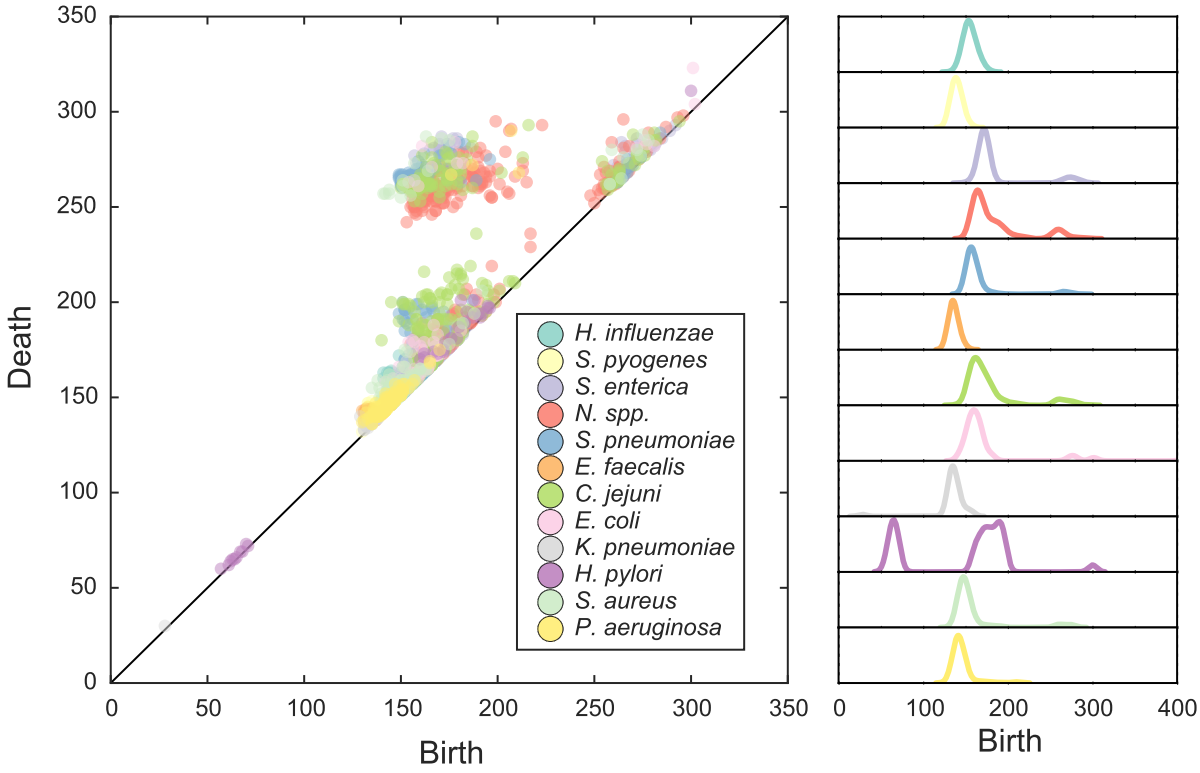


Figure 7.2: The H_1 persistence diagram for the twelve pathogenic strains selected for this study using MLST profile data. There are three broad scales of recombination. To the right is the birth time distribution for each strain. *H. pylori* has an earlier scale of recombination not present in the other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

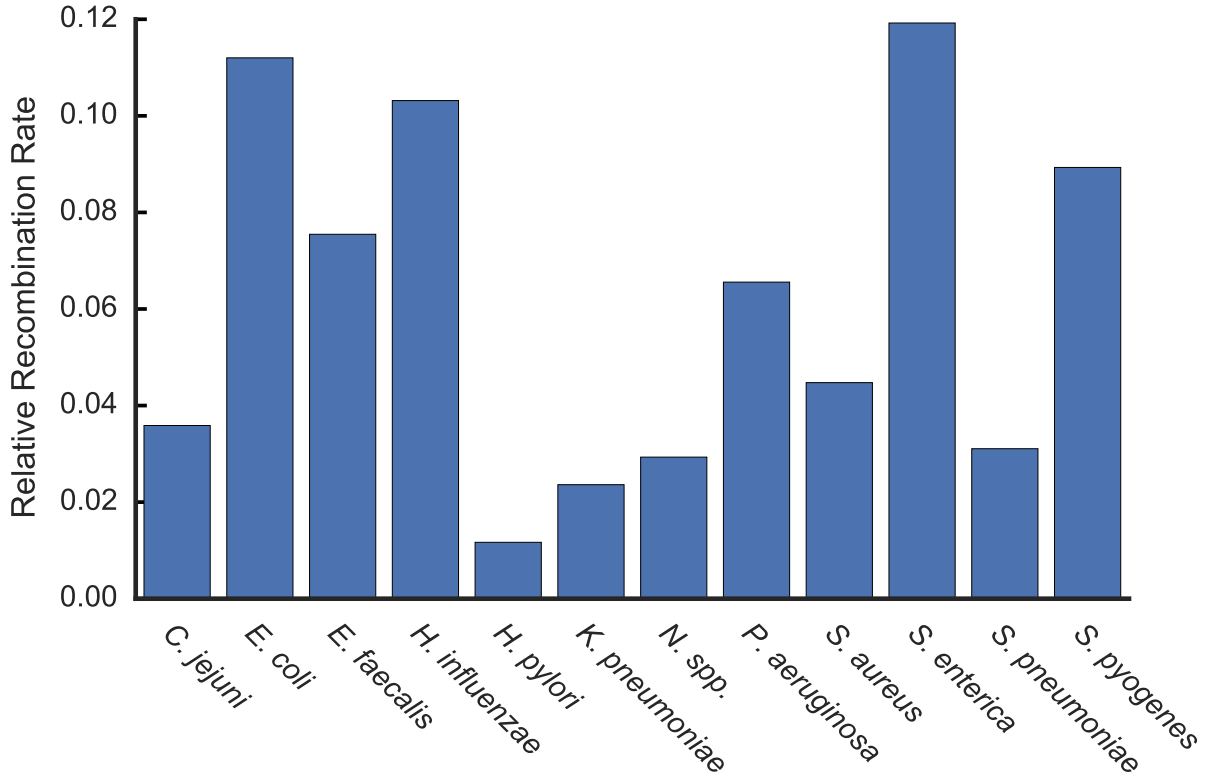


Figure 7.3: Relative recombination rates computed by persistent homology from MLST profile data.

tions Wattam et al., 2013. The FigFam annotation scheme consists of over 100,000 protein families curated from over 950,000 unique proteins Meyer, Overbeek, and Rodriguez, 2009.

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 Fraser, Hanage, and Spratt, 2007. 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

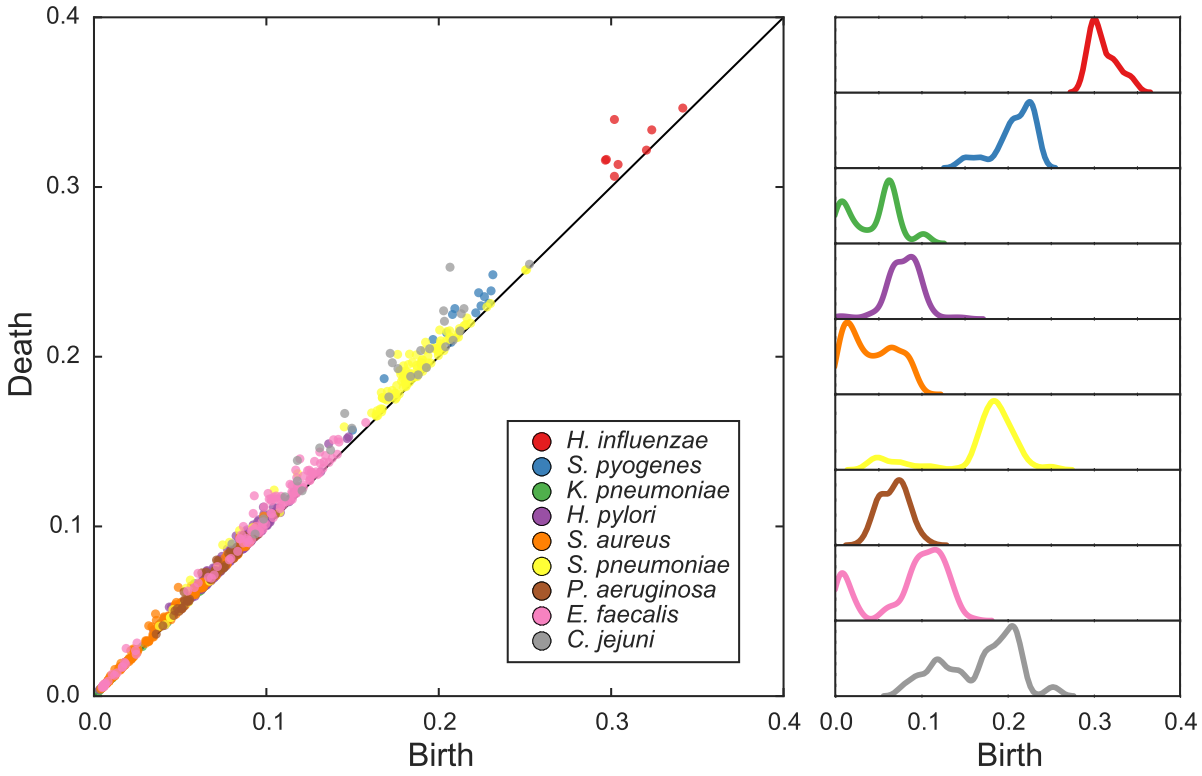


Figure 7.4: Persistence diagram for a subset of pathogenic bacteria, computed using the FigFam annotations compiled in PATRIC. Compared to the MLST persistence diagram, the Figfam diagram has a more diverse scale of topological structure.

recombination scales. The large scales of exchange in *H. influenzae* suggest it can regularly acquire novel genetic material from distantly related strains.

7.4 Antibiotic Resistance in *Staphylococcus aureus*

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

codes for a dysfunctional penicillin-binding protein 2a (PBP2a), which inhibits β-lactam antibiotic binding, the primary mechanism of action Jensen and Lyon, 2009. 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 Inc., 2015. 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 Cooper and Feil, 2006.

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.

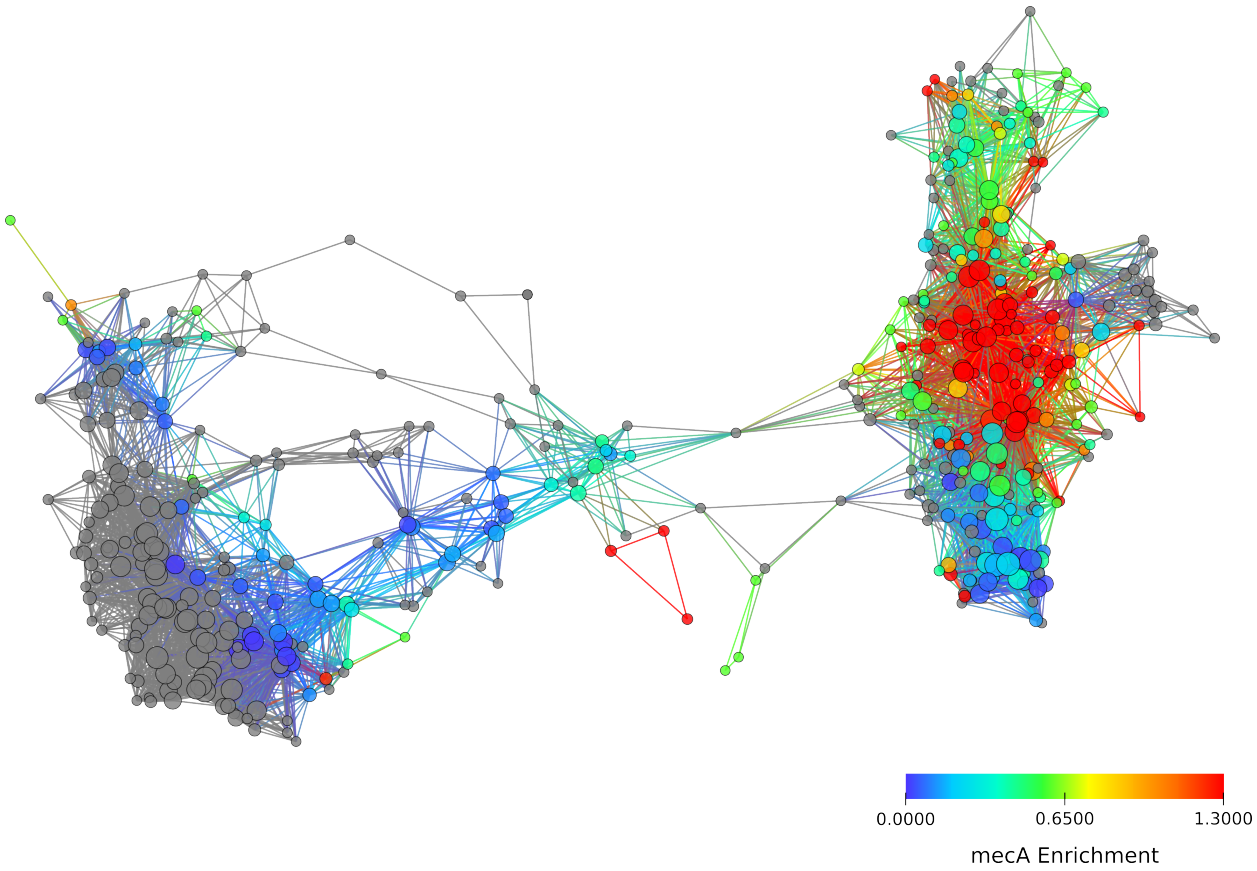


Figure 7.5: The FigFam similarity network of *S. aureus* constructed using Mapper as implemented in Ayasdi Iris. We use a Hamming metric and Primary and Secondary Metric SVD filters (res: 30, gain 4x, eq.). Node color is based on strain enrichment for *mecA*, the gene conferring β-Lactam resistance.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

7.5 Microbiome as a Reservoir of Antibiotic Resistance Genes

While antibiotic resistance can be acquired through gene exchange between strains of the same species, it is also possible for gene exchange to occur between distantly related species. It has been recognized that an individual's microbiome, the set of microorganisms that exist symbiotically within a human host, can act as a reservoir of antimicrobial resistance genes Sommer, Church, and Dantas, 2010; Penders et al., 2013. 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 Consortium, 2012. The HMP has defined a set of reference strains that have been observed in human microbiomes. We collect FigFam annotations from PATRIC for the reference strain list in the gastrointestinal tract. We focus on the gastrointestinal tract because it is an isolated environment and likely to undergo higher rates of exchange than other anatomic regions. Of the 717 reference strains, 321 had FigFam annotations. We computed a similarity matrix as in previous sections, using correlation as distance. The resulting network is shown in Figure 7.6, where strains are colored by phyla-level classifications. While largely recapitulating phylogeny, the network depicts interesting correlations between phya, such as the loop between Firmicutes, Bacteroides, and Proteobacteria.

Next, we searched for genomic annotations relating to β-lactam resistance. 10 strains in the reference set had matching annotations, and we highlight those strains in the network with green diamonds. We observe resistance mostly concentrated in the Firmicutes, of which *S. aureus* is a member, however there is a strain of Proteobacteria that has acquired the resistance gene. Transfer of beta-lactam resistance into the Protebacteria is clinically 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 Consortium, 2012. In this case, an individuals personal microbiome network will differ from the network we show in Figure 7.6, which was constructed from the set of *all* strains that have been reported across studies of multiple individuals. The relative risk for acquiring self-induced resistance will therefore vary from person to person and by the infectious strain acquired. However, a network analysis of this type will give clues as to possible routes by which antibiotic resistance may be acquired. In

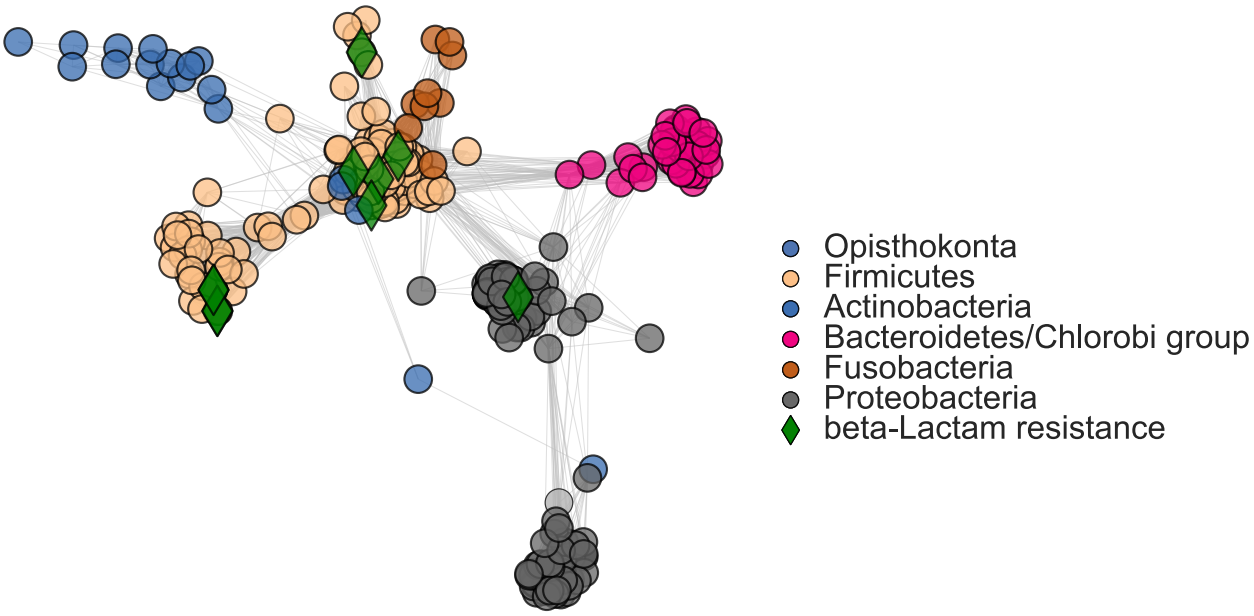


Figure 7.6: The FigFam similarity network of gastrointestinal tract reference strains identified in the Human Microbiome Project. The green diamond identifies the strains carrying resistance to β-Lactam antibiotics.

the clinical setting, this could assist in developing personalized antibiotic treatment regimens. We propose a more thorough expansion of this work, examining the full range of antibiotic resistance genes in order to quantify microbiome risk factors for treatment failure. We foresee an era of genomically informed infectious disease management in the clinical setting, based on an understanding of a patient's personal microbiome profile.

7.6 Conclusions

In this chapter we have used some ideas from topological data analysis to bear on problems in pathogenic microbial genetics. First, we used persistent homology to evaluate recombination rates in the core genome using MLST profile data. We showed that different pathogens have different recombination rates. We expanded this to gene transfer across the whole genome by using protein family annotations in the PATRIC database. We found different scales of recombination in different pathogens. Second, we explored the spread of MRSA in *S. aureus* populations using topological methods. We noted increasing resistance in a

previously isolated population. Finally, we studied the emergence of β-lactam resistance in the microbiome, and proposed methods by which personal risk could be assessed by microbiome typing. These results point to a role for graph mining and topological data mining in health and personalized medicine.

Chapter 8

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 Galperin et al., 2014. 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 Galperin et al., 2014

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.

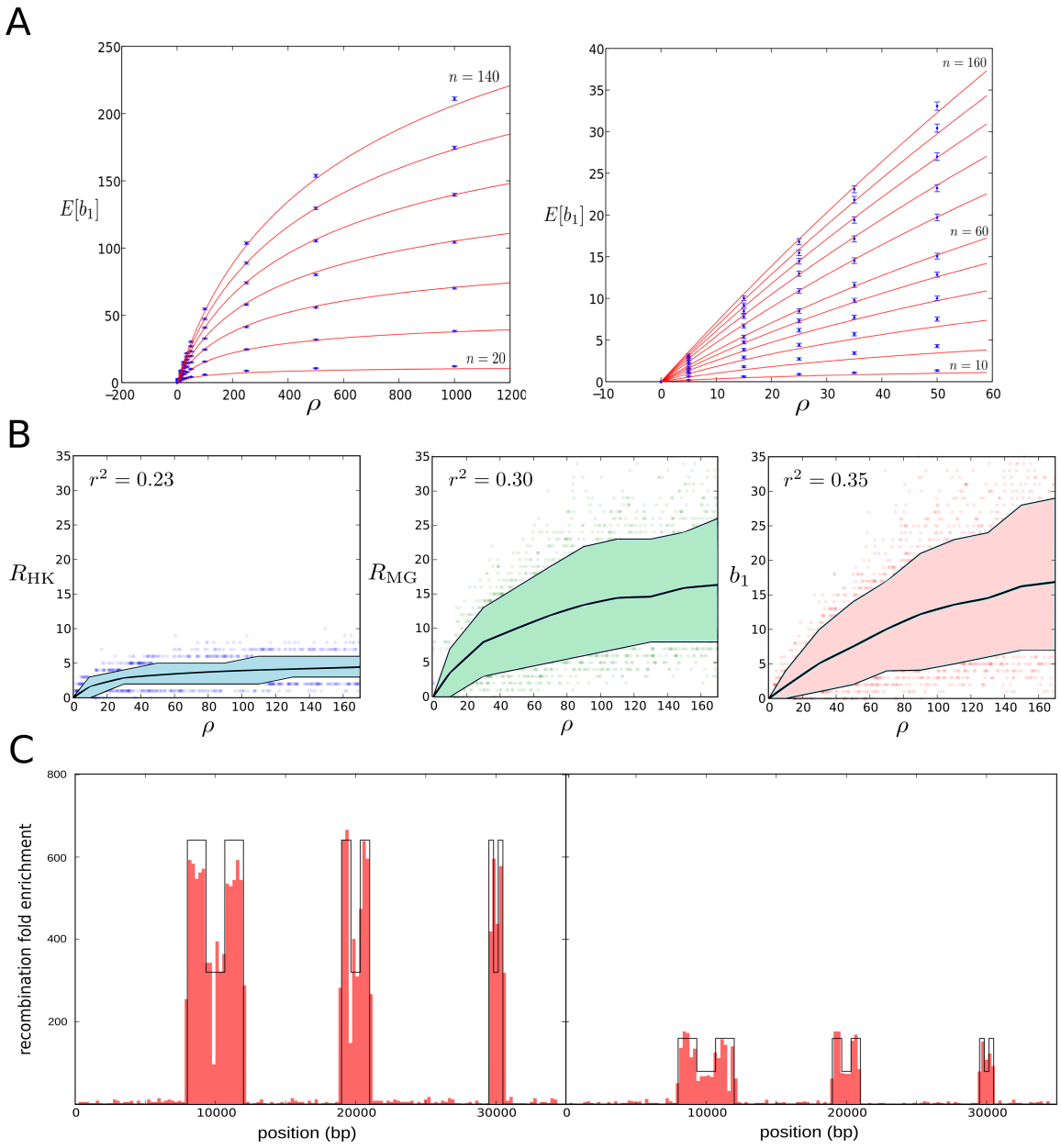


Figure 9.1: Calibration

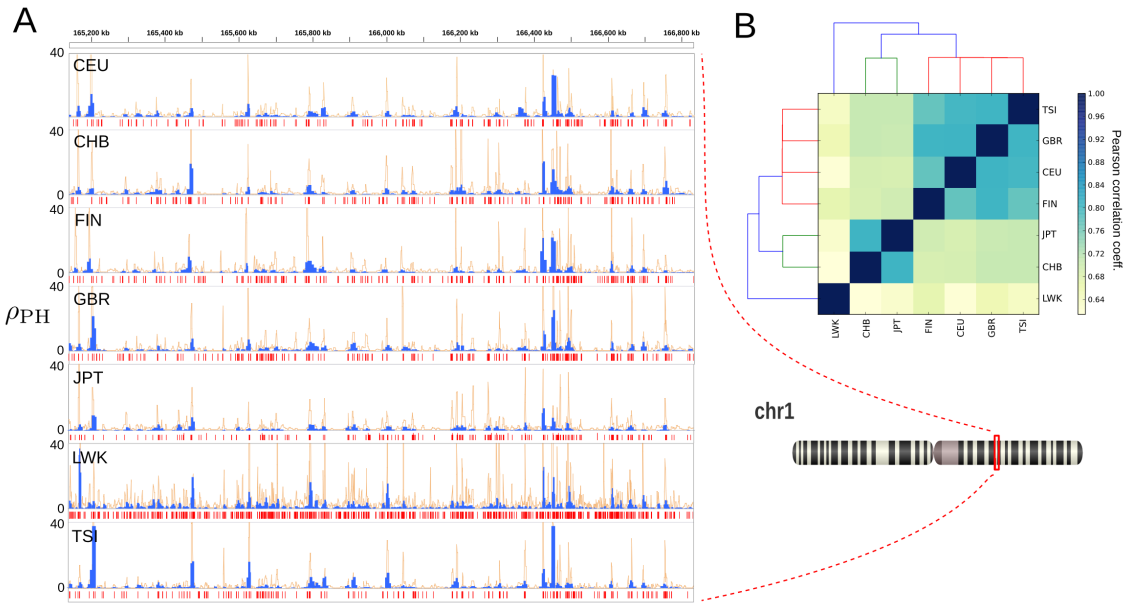


Figure 9.2: Population Tracks

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 Dekker, Marti-Renom, and Mirny, 2013. Chromatin conformation is dynamic, and will change throughout the cellular cycle, under the infuence of a diverse range of chromatin remodeling proteins, such as CTCF. Chromatin architecture can further be controlled epigenetically through post-translational modifications including methylation and phosphorylation.

Recently developed experimental approaches have provided unprecedented high-throughput access into the three dimensional architecture of DNA inside the nucleus Lieberman-Aiden et al., 2009; Dekker, Marti-Renom, and Mirny, 2013; Ay and Noble, 2015. 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 Ay and Noble, 2015. 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 Chan, Carlsson, and Rabadan, 2013; K. J. Emmett and Rabadan, 2014. 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.

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 Dekker, Marti-Renom, and Mirny, 2013. 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 Ay and Noble, 2015 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 Carlsson, 2014 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 MacPherson and Schweinhart, 2012, we define the *size*

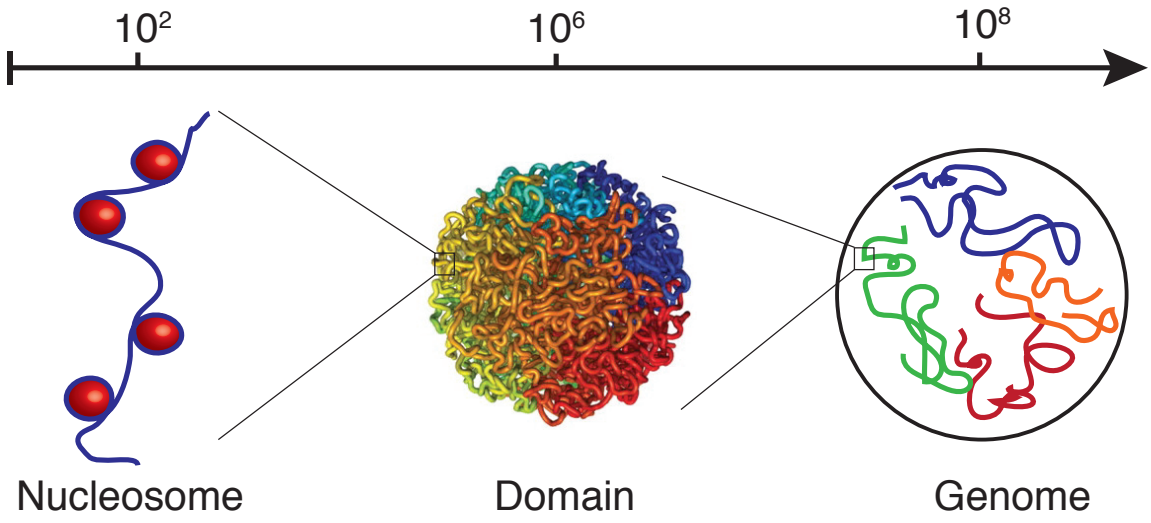


Figure 10.1: Three hierarchices of chromatin organization. At the 100 bp scale, DNA chains wrap around protein complexes called nucleosomes. At the megabase scale, these chains are compacted into domains. Lieberman-Aiden *et al.* proposed closed domains form a *fractal globule* structure. At the genome scale, chromosomes fold into the nucleus in separate territories. The fractal globule represents densely packed regions not open for transcription. Fractal globule image from Lieberman-Aiden et al., 2009. Reprinted with permission from AAAS.

of a PH class as

$$x_i = \frac{b_i + d_i}{2}. \tag{10.1}$$

The distribution of PH class sizes reflects the scales of folding observed. We use Dionysus to compute persistent homology Morozov, 2012.

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

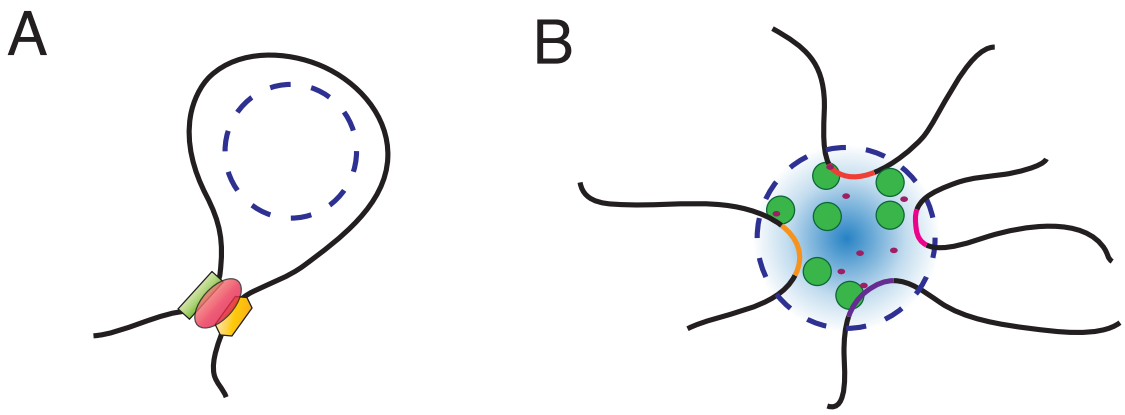


Figure 10.2: Two examples of long range chromatin interactions resulting in a topological loop. (A) A protein mediated (red) point-interaction between an enhancer (green) and promoter (yellow) sequence. (B) A transcription factory consists of dense RNA polymerase (green) around a structural core in which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

of polymerase proteins. This phenomenon is known as a transcription factory, and genes adjacent to a given factory will have correlated levels of expression. We call this interaction a *multi-jump loop*, because the minimal hole generated by the filtration will span multiple nonadajacent loci.

10.2.2 Minimal Cycle Algorithm

It is important to be able to localize a cycle in order to annotate particular loops an interactions. To define a notion of minimal cycle corresponding to a PH class, we first use the contact map to locate an "essential edge" which a cycle must contain. To do this, the values of the heatmap are perturbed so that they are unique and there is a well-defined map from PH birth times to pairs of chromatin segments. That is, we can associate to each PH class (b_i, d_i) a unique "essential edge" that enters the filtration at the birth time b_i. We define a minimal cycle corresponding to (b_i, d_i) to be one containing the essential edge that traverses the shortest length along the genome, and is homologically independent from the

minimal cycles of all classes born before b_i Schweinhart, 2015. This does not uniquely specify a cycle, and we break ties by preferring cycles with shorter jumps. Specifically, if $x\ y\ ...\ z$ is homologous to $x\ x+1\ y\ ...\ z$ (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 Doyle et al., 2014 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 Le et al., 2013. *C. crescentus* has a 4MB circular genome. In that paper, chromatin interaction domains

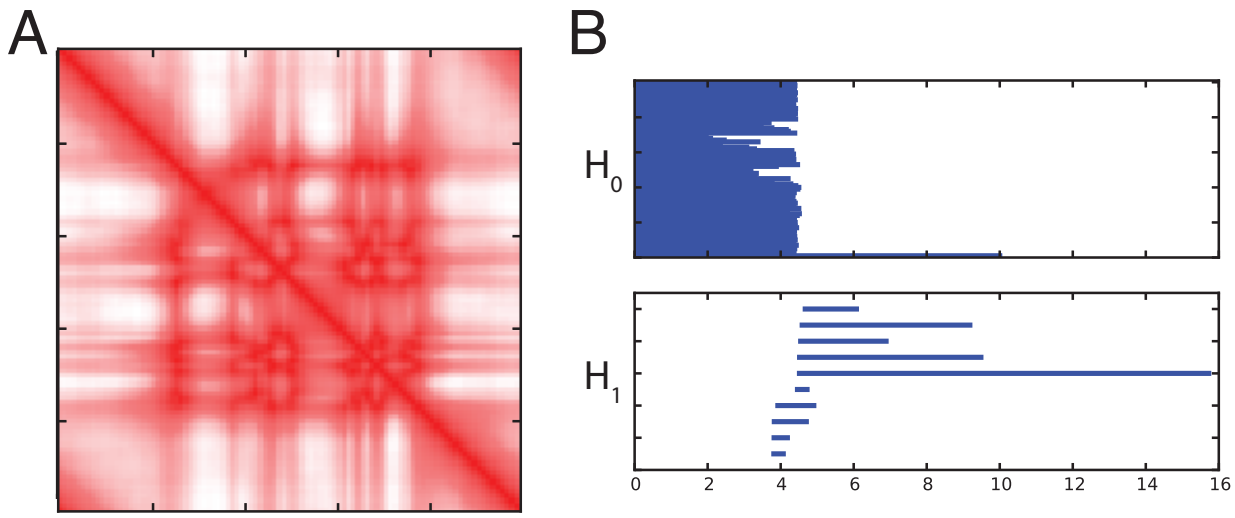


Figure 10.3: Polymer Simulation. (A) 50 Mb polymer with 10 fixed loops is allowed to reach an equilbrium conformation. (B) PH identifies 10 H_1 loops.

(CIDs) were identified at scales between 30 to 420 kb. The authors proposed a structural model consisting of brush-like plectonemes arranged along the circular fiber.

Here we look at look at sample GSM1120446, a wildtype *Caulobacter* cell. In Figure 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 10.4C 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 Wang, Llopis, and Rudner, 2013), 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 Le et al., 2013, but rather reflect larger scale folding patterns.

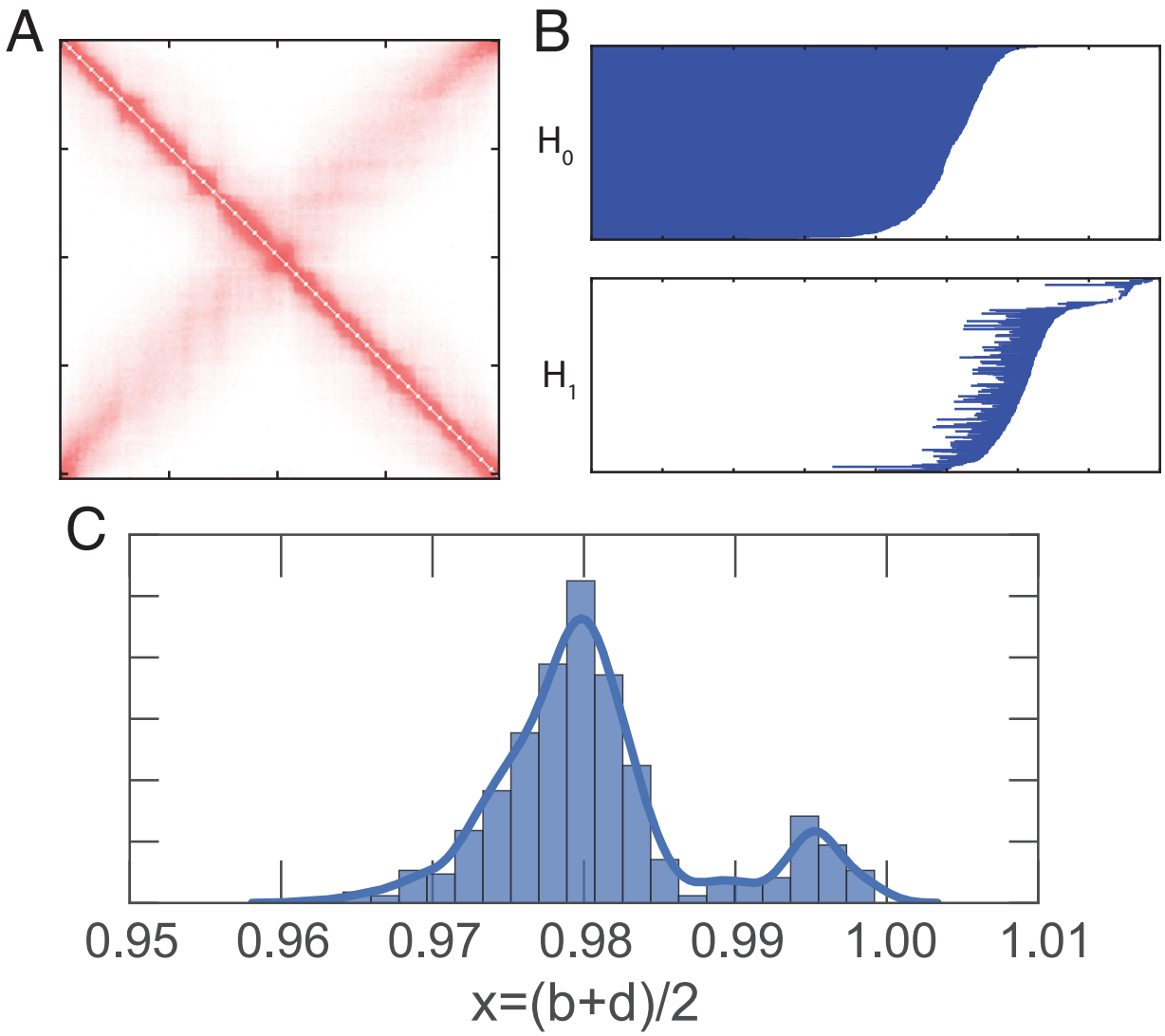


Figure 10.4: (A) Contact map and (B) barcode diagram for *Caulobacter*. (C) Distribution of H_1 bar sizes for *Caulobacter* shows a bimodal scale of folding patterns.

10.5 Human Data

We examined one of the original human Hi-C data sets as published in Lieberman-Aiden et al., 2009. 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 the megabase scale, they proposed a fractal globule model in which nearby loci along the polymer are spatially proximate in 3D.

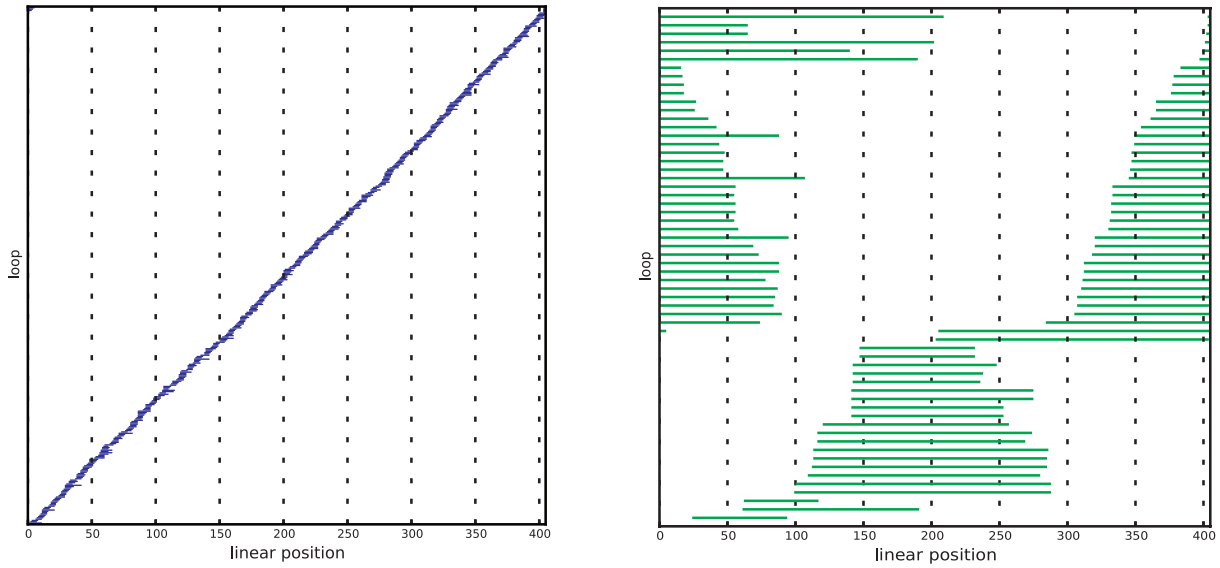


Figure 10.5: Minimal cycles projected linearly for *Caulobacter*. Left: small loops distribute uniformly across the genome. Right: pattern of large loops, which segregate into two chromosomal domains.

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.

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 Lieberman-Aiden et al., 2009, 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 (200bp). More recent work has yielded Hi-C datasets at kilobase resolution Jin et al., 2013; Rao et al., 2014, 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.

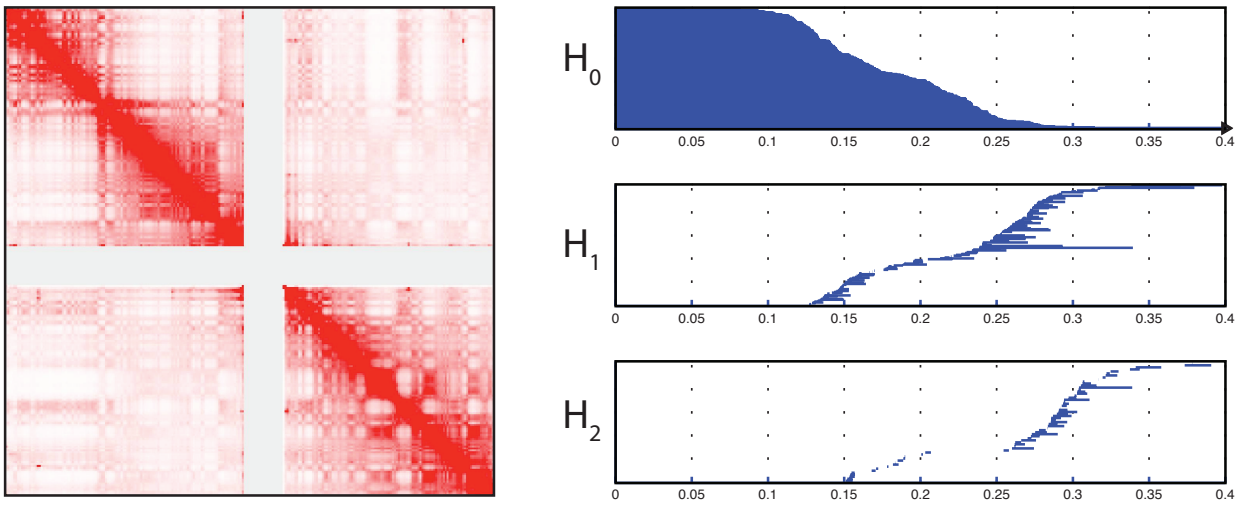


Figure 10.6: Hi-C data chromosome 1 from GM06690 human cell line data, from Lieberman-Aiden et al., 2009. Left: Contact map representation. Right: PH identifies complex multi-scale topology.

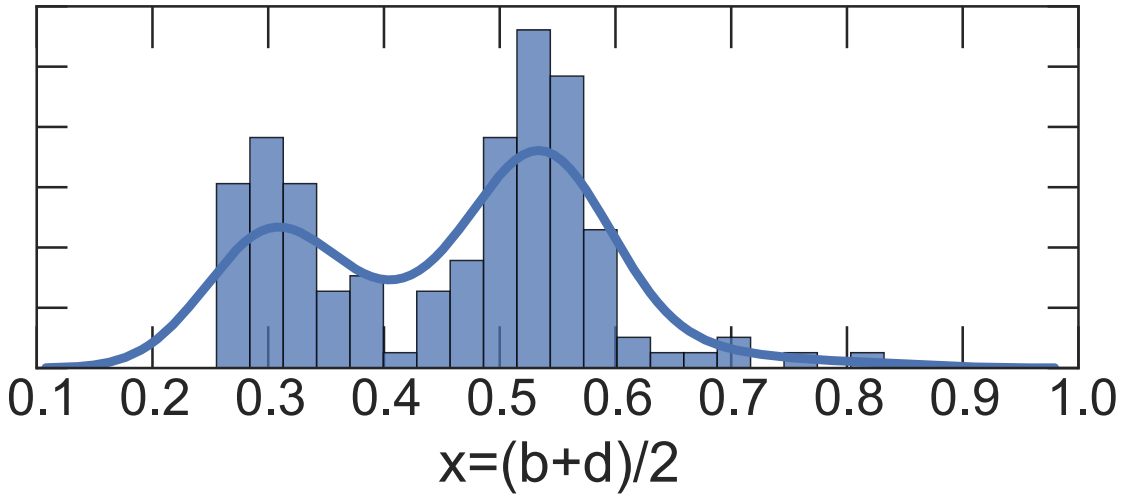


Figure 10.7: Distribution of H_1 bar sizes for GM06690 human cell line data shows a bimodal scale of folding patterns.

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

Chapter 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] Robert Adler et al. “Persistent Homology for Random Fields and Complexes”. In: *arXiv.org* (2010). arXiv: 1003.1001.

[2] Ferhat Ay and William S Noble. “Analysis methods for studying the 3D architecture of the genome”. In: *Genome Biology* 16.1 (2015), p. 1306.

[3] Hans-Jürgen Bandelt and Andreas WM Dress. “A canonical decomposition theory for metrics on a finite set”. In: *Advances in Mathematics* 92.1 (1992).

[4] Hans-Jürgen Bandelt, Peter Forster, and Arne Röhl. “Median-joining networks for inferring intraspecific phylogenies.” In: *Molecular Biology and Evolution* 16.1 (1999), pp. 37–48.

[5] L.J. Billera, S.P. Holmes, and K. Vogtmann. “Geometry of the space of phylogenetic trees”. In: *Advances in Applied Mathematics* 27.4 (2001), pp. 733–767.

[6] Andrew J Blumberg et al. “Robust Statistics, Hypothesis Testing, and Confidence Intervals for Persistent Homology on Metric Measure Spaces”. In: *Foundations of Computational Mathematics* 14.4 (2014), pp. 745–789.

[7] Karol Borsuk. “On the imbedding of systems of compacta in simplicial complexes”. In: *Fundamenta Mathematicae* 35.1 (1948), pp. 217–234.

[8] Peter J Bowler. *Evolution: The History of an Idea*. University of California Press, 2003.

[9] Peter Bubenik. “Statistical Topological Data Analysis using Persistence Landscapes”. In: *Journal of Machine Learning Research* 16 (2015), pp. 77–102.

[10] Peter Bubenik and Peter T Kim. “A statistical approach to persistent homology”. In: *Homology, Homotopy and Applications* 9.2 (2007), pp. 337–362.

[11] Gunnar Carlsson. “Topology and data”. In: *Bulletin of the American Mathematical Society* 46.2 (2009), pp. 255–308.

[12] Gunnar Carlsson. "Topological pattern recognition for point cloud data". In: *Acta Numerica* 23 (2014), pp. 289–368.

[13] Gunnar Carlsson and Afra Zomorodian. "The theory of multidimensional persistence". In: *Discrete & Computational Geometry* 42.1 (2009), pp. 71–93.

[14] Luca L Cavalli-Sforza and Anthony WF Edwards. "Phylogenetic analysis. Models and estimation procedures". In: *American Journal of Human Genetics* 19.3 (1967), pp. 550–570.

[15] Joseph Chan, Gunnar Carlsson, and Raul Rabadan. "Topology of Viral Evolution". In: *Proceedings of the National Academy of Sciences* 110.46 (Nov. 2013), pp. 18566–18571. DOI: 10.1073/pnas.1313480110.

[16] Frédéric Chazal, David Cohen Steiner, et al. "GromovHausdorff Stable Signatures for Shapes using Persistence". In: *Computer Graphics Forum*. Wiley Online Library, 2009, pp. 1393–1403.

[17] Frédéric Chazal, Marc Glisse, et al. "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.

[18] Francesca D Ciccarelli et al. "Toward Automatic Reconstruction of a Highly Resolved Tree of Life". In: *Science* 311.5765 (2006), pp. 1283–1287.

[19] The Human Microbiome Project Consortium. "Structure, function and diversity of the healthy human microbiome". In: *Nature* 486.7402 (2012), pp. 207–214.

[20] Jessica E Cooper and Edward J Feil. "The phylogeny of *Staphylococcus aureus* - which genes make the best intra-species markers?" In: *Microbiology* 152.5 (2006), pp. 1297–1305.

[21] Francis Crick. "Central dogma of molecular biology". In: *Nature* 227 (1970), pp. 561–563.

[22] Tal Dagan and William Martin. "The tree of one percent". In: *Genome Biology* 7.10 (2006), p. 118.

[23] Charles Darwin. *On the Origin of Species by Means of Natural Selection*. London: Murray, 1859.

[24] Vin de Silva and Gunnar Carlsson. "Topological estimation using witness complexes". In: *Proceedings of the First Eurographics conference on Point-Based Graphics*. Eurographics Association. 2004, pp. 157–166.

[25] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data". In: *Nature Reviews Genetics* 14 (2013), pp. 390–403.

[26] William Ford Doolittle. "Phylogenetic classification and the universal tree". In: *Science* 284.5423 (1999), pp. 2124–2128. DOI: 10.1126/science.284.5423.2124.

[27] William Ford Doolittle and R Thane Papke. "Genomics and the bacterial species problem". In: *Genome Biology* 7.9 (2006), p. 116. DOI: 10.1186/gb-2006-7-9-116.

[28] Boryana Doyle et al. "Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions". In: *PLoS Computational Biology* 10.10 (2014), e1003867.

[29] Andreas Dress, Katharina Huber, and Vincent Moulton. "Some variations on a theme by Buneman". In: *Annals of Combinatorics* 1.1 (1997), pp. 339–352.

[30] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[31] Kevin J Emmett and Raul Rabadan. "Characterizing Scales of Genetic Recombination and Antibiotic Resistance in Pathogenic Bacteria Using Topological Data Analysis". In: *Brain Informatics and Health*. Ed. by Dominik Slezak et al. Vol. 8609. Lecture Notes in Computer Science. Springer, 2014, pp. 540–551.

[32] Kevin Emmett and Raul 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.

[33] Kevin Emmett, Daniel Rosenbloom, et al. "Parametric Inference using Persistence Diagrams: A Case Study in Population Genetics". In: *ICML Workshop on Topological Methods in Machine Learning*. 2014.

[34] Brittany Terese Fasy et al. "Confidence sets for persistence diagrams". In: *Ann. Statist.* 42.6 (), pp. 2301–2339. DOI: 10.1214/14-AOS1252.

[35] Walter M Fitch and Emanuel Margoliash. "Construction of Phylogenetic Trees". In: *Science* 3760 (1967), pp. 279–284. DOI: 10.1126/science.155.3760.279.

[36] Christophe Fraser, William P Hanage, and Brian G Spratt. "Recombination and the Nature of Bacterial Speciation". In: *Science* 315.5811 (2007), pp. 476–480.

[37] Michael Y Galperin et al. "Expanded microbial genome coverage and improved protein family annotation in the COG database". In: *Nucleic Acids Research* (43 2014), pp. D261–D269.

[38] Bernd Gärtner. "Fast and robust smallest enclosing balls". In: *Algorithms-ESA99*. Springer, 1999, pp. 325–338.

[39] Robert Ghrist. "Barcodes: The persistent topology of data". In: *Bulletin of the American Mathematical Society* 45.1 (2008), pp. 61–75.

[40] Galina Glazko et al. "Evolutionary history of bacteriophages with double-stranded DNA genomes". In: *Biology Direct* 2.1 (2007), p. 36.

[41] Stephen Jay Gould. "The Structure of Evolutionary Theory". In: (2002).

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

[43] Allen Hatcher. *Algebraic Topology*. Cambrdige University Press, 2002.

[44] Richard R Hudson. "Generating samples under a Wright–Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2 (2002).

[45] Richard R Hudson and Norman L Kaplan. "Statistical properties of the number of recombination events in the history of a sample of DNA sequences". In: *Genetics* 111.1 (1985), pp. 147–164.

[46] Daniel H Huson, Regula Rupp, and Céline Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2010.

[47] Julian Huxley. *Evolution: The Modern Synthesis*. MIT Press, 1942.

[48] Ayasdi Inc. *Iris*. http://www.ayasdi.com. 2015.

[49] Slade O Jensen and Bruce R Lyon. "Genetics of antimicrobial resistance in *Staphylococcus aureus*". In: *Future Microbiology* 4.5 (2009), pp. 565–582.

[50] Fulai Jin et al. "A high-resolution map of the three-dimensional chromatin interactome in human cells". In: *Nature* (2013).

[51] Keith A Jolley and Martin CJ Maiden. "BIGSdb: Scalable analysis of bacterial genome variation at the population level". In: *BMC Bioinformatics* 11.1 (2010), p. 595.

[52] Tomasz Kaczynski, Konstantin Mischaikow, and Marian Mrozek. *Computational Homology*. Vol. 157. Applied Mathematical Sciences. Springer, Jan. 2004.

[53] Motoo Kimura. "Evolutionary rate at the molecular level". In: *Nature Reviews Genetics* 217.5129 (1968), p. 624.

[54] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1984.

[55] Martin Kreitman. "Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster". In: *Nature* 304.5925 (Aug. 1983), pp. 412–417.

[56] David M Kristensen et al. "Orthologous gene clusters and taxon signature genes for viruses of prokaryotes". In: *Journal of Bacteriology* 195.5 (2013), pp. 941–950.

[57] Jeffrey G Lawrence, Graham F Hatfull, and Roger W Hendrix. "Imbroglios of Viral Taxonomy: Genetic Exchange and Failings of Phenetic Approaches". In: *Journal of Bacteriology* 184.17 (2002), pp. 4891–4905.

[58] Tung BK Le et al. "High-resolution mapping of the spatial organization of a bacterial chromosome". In: *Science* 342.6159 (2013), pp. 731–734.

[59] Michael Phillip Lesnick. "Multidimensional Interleavings and Applications to Topological Inference". PhD thesis. Stanford University, 2012.

[60] E Lieberman-Aiden et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome". In: *Science* 326.5950 (2009), pp. 289–293.

[61] G Lima-Mendez et al. "Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes". In: *Molecular Biology and Evolution* 25.4 (2008), pp. 762–777.

[62] P Y Lum et al. "Extracting insights from the shape of complex data using topology". In: *Scientific Reports* 3 (Feb. 2013).

[63] Robert MacPherson and Benjamin Schweinhart. "Measuring shape with topology". In: *Journal of Mathematical Physics* 53.7 (2012), p. 073516.

[64] F Meyer, R Overbeek, and A Rodriguez. "FIGfams: yet another set of protein families". In: *Nucleic Acids Research* 37.20 (2009), pp. 6643–6654.

[65] Yuriy Mileyko, Sayan Mukherjee, and John Harer. "Probability measures on the space of persistence diagrams". In: *Inverse Problems* 27.12 (2011), p. 124007.

[66] Dmitriy Morozov. *Dionysus: a C++ library for computing persistent homology*. 2012.

[67] Daniel Müllner and Aravindakshan Babu. *Python Mapper: An open-source toolchain for data exploration, analysis and visualization*. 2013.

[68] Harold C Neu. "The Crisis in Antibiotic Resistance". In: *Science* 257.5073 (1992), pp. 1064–1073.

[69] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival". In: *Proceedings of the National Academy of Sciences of the United States of America* 108.17 (2011), pp. 7265–7270.

[70] Howard Ochman, Jeffrey G Lawrence, and Eduardo A Groisman. "Lateral gene transfer and the nature of bacterial innovation". In: *Nature* 405.6784 (2000), pp. 299–304.

[71] Maureen A O'Malley and Eugene V Koonin. "How stands the Tree of Life a century and a half after The Origin?" In: *Biology Direct* 6.1 (2011), pp. 1–21.

[72] World Health Organization. *Antimicrobial Resistance: global report on surveillance 2014*. Tech. rep. 2014.

[73] John Penders et al. "The human microbiome as a reservoir of antimicrobial resistance". In: *Frontiers in microbiology* 4 (2013).

[74] Suhas S P Rao et al. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". In: *Cell* (2014), pp. 1–16.

[75] Holger Rohde et al. "Open-Source Genomic Analysis of Shiga-Toxin–Producing *E. coli* O104:H4". In: *New England Journal of Medicine* 365.8 (2011), pp. 718–724.

[76] Forest Rohwer and Rob Edwards. "The Phage Proteomic Tree: a genome-based taxonomy for phage". In: *Journal of Bacteriology* 184.16 (2002), pp. 4529–4535.

[77] Naruya Saitou and Masatoshi Nei. "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees". In: *Molecular Biology and Evolution* 4.4 (1987), pp. 406–425.

[78] Benjamin Schweinhart. "Statistical Topology of Embedded Graphs". PhD thesis. Princeton University Press, 2015.

[79] Gurjeet Singh, Facundo Mémoli, and Gunnar 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.

[80] Morten O Sommer, George M Church, and Gautam Dantas. "The human microbiome harbors a diverse reservoir of antibiotic resistance genes". In: *Virulence* 1.4 (2010), pp. 299–303.

[81] Yun S Song and Jotun Hein. "Constructing minimal ancestral recombination graphs". In: *Journal of Computational Biology* 12.2 (2005), pp. 147–169. DOI: 10.1089/cmb.2005.12.147.

[82] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. "JavaPlex: A research software package for persistent (co)homology". In: *Proceedings of ICMS 2014*. Ed. by Han Hong and Chee Yap. Lecture Notes in Computer Science 8592. 2014, pp. 129–136.

[83] International Committee on Taxonomy of Viruses. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. Ed. by Andrew M Q King et al. Immunology and Microbiology 2011. Academic Press, 2012.

[84] Christopher M Thomas and Kaare M Nielsen. "Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria". In: *Nature Reviews Microbiology* 3.9 (2005), pp. 711–721.

[85] Katharine Turner et al. "Fréchet Means for Distributions of Persistence diagrams". In: *arXiv.org* (2012). arXiv: 1206.2790v2 [math.ST].

[86] John Wakeley. *Coalescent Theory*. Roberts & Company, 2009.

[87] Xindan Wang, Paula Montero Llopis, and David Z Rudner. "Organization and segregation of bacterial chromosomes". In: *Nature Reviews Genetics* 14.3 (2013), pp. 191–203.

[88] James D Watson and Francis HC Crick. "Molecular structure of nucleic acids". In: *Nature* 171.4356 (1953), pp. 737–738.

[89] A R Wattam et al. "PATRIC, the bacterial bioinformatics database and analysis resource". In: *Nucleic Acids Research* 42.D1 (2013), pp. D581–D591.

[90] Carl R Woese and George E Fox. "Phylogenetic structure of the prokaryotic domain: the primary kingdoms". In: *Proceedings of the National Academy of Sciences of the United States of America* 74.11 (1977), pp. 5088–5090.

[91] Afra Zomorodian. *Topology for Computing*. Cambridge University Press, 2005.

[92] Emile Zuckerkandl and Linus Pauling. "Molecular disease, evolution, and genetic heterogeneity". In: *Horizons in Biochemistry*. Ed. by M Kasha and B Pullman. Academic Press, 1962, pp. 189–225.

[93] Emile Zuckerkandl and Linus Pauling. "Molecules as documents of evolutionary history". In: *Journal of Theoretical Biology* 8.2 (1965), pp. 357–366.